IJCSIS Vol. 15 No. 12 (II), December 2017
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2017 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

*It is our great pleasure to present the **December 2017 issue** (Volume 15 Number 12) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and technology. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 9800 times and this journal is experiencing steady and healthy growth. Google statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, Academia.edu and EBSCO among others.*

A great journal cannot be made great without a dedicated editorial team of editors and reviewers. On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status, making sure we deliver high-quality content to our readers in a timely fashion.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." We would like to thank you, the authors and readers, the content providers and consumers, who have made this journal the best possible.

*For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 15, No. 12, December 2017 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	**Dr Riktesh Srivastava [Profile]** Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	**Dr. Jianguo Ding [Profile]** Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran	**Dr. Naseer Alquraishi [Profile]** University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	**Dr. Kai Cong [Profile]** Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	**Dr. Omar A. Alzubi [Profile]** Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	**Dr. Jorge A. Ruiz-Vanoye [Profile]** Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	**Prof. Ning Xu,** Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	**Dr . Bilal Alatas [Profile]** Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	**Dr. Ioannis V. Koskosas,** University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	**Dr Venu Kuthadi [Profile]** University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	**Dr. Zhihan lv [Profile]** Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	**Prof. Ghulam Qasim [Profile]** University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	**Prof. Dr. Maqbool Uddin Shaikh [Profile]** Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	**Dr. Musa Peker [Profile]** Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	**Dr. Wencan Luo [Profile]** University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	**Dr. Ijaz Ali Shoukat [Profile]** King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	**Dr. Yilun Shang [Profile]** Tongji University, Shanghai, China
Dr. Sedat Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

6. PaperID 30111706: Robotics and Artificial Intelligence: Differences and Similarities (pp. 26-28)

(1) Omankwu, Obinnaya Chinecherem; (2) Nwagu, Chikezie Kenneth; (3) Inyiama, Hycient
(1) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(2) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

7. PaperID 30111707: Faculty Assignment and Timetabling using Optimisation (pp. 29-35)

MA Nang Laik, Senior Lecturer, School of Business, Singapore University of Social Sciences, Singapore
SEN Prabir, Director Research, STATGRAF Research, Vancouver, Canada

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

8. PaperID 30111709: A Novel and Robust Retroareolar Region Detection in Mammography Images (pp. 36-39)

Dr. P. Pedda Sadhu Naik, Professor & HOD, Dr. Samuel George Institute of Engineering & Technology, Markapur, AP.
Dr T. Venugopal, Professor & HOD, JNTUH College of Engineering, Jagityala, Telangana

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

9. PaperID 30111712: Improving DBMS Security through the use of a Checksum Technique (pp. 40-48)

*Naim Ajlouni *, Firas Ajlouni **, Alaa Ali Hameed ****
*Istanbul Aydin University * ****
*Lancashire College of Further Education ***

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

10. PaperID 30111713: Coupling GIS-based MCA and AHP techniques for Hospital Site Selection (pp. 49-56)

Aseel B. Kmail, Jacqueleen Jubran, Walid Sabbah
Geographic Information System Department, AAUJ, Jenin, Palestine

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

11. PaperID 30111714: Big and Connected Data Analysis with Graph and Relational Databases Using Collaborative Filtering Technique (pp. 57-65)

Joseph George Davis, Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
Mr Joseph K. Panford, Senior Lecturer, Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

(1) Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
(2) Computer Science Department, Michael Okpara University of Agriculture Umudike Umuahia, Abia State, Nigeria.
(3) Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

23. PaperID 30111741: Credit Card Fraud Detection Using Neural Network (pp. 176-180)

(1) Nwagu, Chikezie Kenneth; (2) Omankwu, Obinnaya Chinecherem; (3) Inyiama, Hycient
(1) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(2) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

24. PaperID 30111742: Reinforcement Learning a Tool for Filtering Personalized Web Document (pp. 181-185)

(1) Nwagu, Chikezie Kenneth; (2) Omankwu, Obinnaya Chinecherem; (3) Inyiama, Hycient
(1) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(2) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

25. PaperID 30111744: Influence of Distributed Generation on the Protection Scheme of Local Distribution System (pp. 186-190)

Usman Yousaf, Dr. Azzam ul Asar, Alishpa Khattak
Department of Electrical Engineering, CECOS University, Peshawar, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

26. PaperID 30111747: Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm (pp. 191-200)

Abdeljalil EL ABDOULI, Larbi HASSOUNI, Houda ANOUN
RITM Laboratory, CED Engineering Sciences, Ecole Superieure de Technologie, Hassan II University of Casablanca, Morocco

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

27. PaperID 30111749: Using Biometric Watermarking for Video File Protection Based on Chaotic Principle (pp. 201-206)

Saja J. Mohammed, Computer Science Department, College of Computer Science & Mathematics, Mosul University Mosul, IRAQ

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

Dr Narasimha Rao Yamarthi (1), Mr Adinew Belay (2), Mr Abdisa Lechisa (3), P.Narasimha Rao (4).
(1) Professor, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
(2) HOD, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
(3) Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
(4) Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

41. PaperID 30111773: An Efficient Voice Based Person Identification System for Secured Communication (pp. 308-325)

Dr. Shakila Basheer, Lecturer in Information system Department, King Khalid University, ABHA, Saudi Arabia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

42. PaperID 30111766: Comparative Analysis of Modern Methods and Algorithms of Cryptographic Protection of Information (pp. 326-330)

Saleh I. Alomar & Saleh A. Khawatreh
AL-Ahliyya Amman University, Department of Engineering

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

43. PaperID 30111775: SQS: An Ontology Based Framework for Evaluating Service Oriented Software Applications: A case study of E-Governance Services (pp. 331-336)

Ezekiel U, Okike, University of Botswana, Gaborone, Botswana

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

Credit Card Fraud Detection Using Neural Network

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem; and Inyiama, Hycient

[1] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[2] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

The payment card industry has grown rapidly the last few years. Companies and institutions move parts of their business, or the entire business, towards online services providing e-commerce, information and communication services for the purpose of allowing their customers better efficiency and accessibility. Regardless of location, consumers can make the same purchases as they previously did "over the desk". The evolution is a big step forward for the efficiency, accessibility and profitability point of view but it also has some drawbacks. The evolution is accompanied with a greater vulnerability to threats. The problem with making business through the Internet lies in the fact that neither the card nor the cardholder needs to be present at the point-of-sale. It is therefore impossible for the merchant to check whether the customer is the genuine cardholder or not. Payment card fraud has become a serious problem throughout the world. Companies and institutions loose huge amounts annually due to fraud and fraudsters continuously seek new ways to commit illegal actions. The good news is that fraud tends to be perpetrated to certain patterns and that it is possible to detect such patterns, and hence fraud. In this paper we will try to detect fraudulent transaction through the neural network along with the genetic algorithm. As we will see that artificial neural network when trained properly can work as a human brain, though it is impossible for the artificial neural network to imitate the human brain to the extent at which brain work, yet neural network and brain, depend for there working on the neurons, which is the small functional unit in brain as well as ANN. Genetic algorithm are used for making the decision about the network topology, number of hidden layers, number of nodes that will be used in the design of neural network for our problem of credit card fraud detection. For the learning purpose of artificial neural network we will use supervised learning feed forward back propagation algorithm.

1. INTRODUCTION

There are many ways in which fraudsters execute a credit card fraud. As technology changes, so does the technology of fraudsters, and thus the way in which they go about carrying out fraudulent activities. Frauds can be broadly classified into three categories, i.e., traditional card related frauds, merchant related frauds and Internet frauds. The different types of methods for committing credit card frauds are described below.

Merchant Related Frauds

Merchant related frauds are initiated either by owners of the merchant establishment or their employees. The types of frauds initiated by merchants are described below:

i. Merchant Collusion **:** This type of fraud occurs when merchant owners or their employees conspire to commit fraud using the cardholder accounts or by using the personal information. They pass on the information about cardholders to fraudsters.

ii. Triangulation: Triangulation is a type of fraud which is done and operates from a web site. The products or goods are offered at heavily discounted rates and are also shipped before payment. The customer while browse the site and if he likes the product he place the online information such as name, address and valid credit card details to the site. When the fraudsters receive these details, they order goods from a legitimate site using stolen credit card details. The fraudsters then by using the credit card information purchase the products.

Internet Related Frauds

The internet is the base for the fraudsters to make the frauds in the simply and the easiest way. Fraudsters have recently begun to operate on a truly transnational level. With the expansion of trans-border, economic and political spaces, the internet has become a new worlds market, capturing consumers from most countries

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

around the world. The below described are most commonly used techniques in Internet fraud:

i. Site clonin: Site cloning is where fraudsters close an entire site or just the pages from which the customer made a purchase. Customers have no reason to believe they are not dealing with the company that they wished to purchase goods or services from because the pages that they are viewing are identical to those of the real site. The cloned site will receive these details and send the customer a receipt of the transaction through the email just as the real company would do. The consumer suspects nothing, while the fraudsters have all the details they need to commit credit card fraud.

ii. *False merchant sites***:** Some sites often offer a cheap service for the customers. That site requests the customer to fill his complete details such as name and address to access the webpage where the customer gets his required products. Many of these sites claim to be free, but require a valid credit card number to verify an individual s age. These kinds of sites in this way collect as many as credit card details. The sites themselves never charge individuals for the services they provide. The sites are usually part of a larger criminal network that either uses the details it collects to raise revenues or sells valid credit card details to small fraudsters.

iii. *Credit card generators***:** These are the computer programs that generate valid credit card numbers and expiry dates. These generators work by generating lists of credit card account numbers from a single account number. The software works by using the mathematical Luhn algorithm that card issuers use to generate other valid card number combinations. This makes the user to allow to illegally generating as many numbers as he desires, in the form of any of the credit card formats.

FRAUD DETECTION USING NEURAL NETWORK

Although there are several fraud detection technology exist based on Data mining, Knowledge Discovery and Expert System etc. but all these are not capable enough to detect the fraud at the time when fraudulent transaction are in progress due to very less chance of a transaction being fraudulent .It has been seen that Credit card fraud detection has two highly peculiar characteristics The first one is obviously the very limited time span in which the acceptance or rejection decision has to be made.

The second one is the huge amount of credit card operations that have to be processed at a given time. To just give a medium size example, millions of Visa card operations take place in a given day, 98% of them being handled on line. Of course, just very few will be fraudulent (otherwise, the entire industry would have soon ended up being out of businesses), but this just means that the haystack where these needles are to be found is simply enormous.

Working principal (Pattern Recognition)

Neural network based fraud detection is based totally on the human brain working principal. Neural network technology has made a computer capable of think. As human brain learn through past experience and use its knowledge or experience in making the decision in daily life problem the same technique is applied with the credit card fraud detection technology. When a particular consumer uses its credit card , there is a fix pattern of credit card use , made by the way consumer uses its credit card.

Using the last one or two year data neural network is train about the particular pattern of using a credit card by a particular consumer. As shown in the figure the neural network are train on information regarding to various categories about the card holder such as occupation of the card holder, income, occupation may fall in one category, while in another category information about the large amount of purchased are placed, these information include the number of large purchase, frequencies of large purchase, location where these kind of purchase are take place etc. within a fixed time period.

In spite of pattern of credit card use neural network are also trained about the various credit card fraud face by a particular bank previously. Based on the pattern of uses of credit card, neural network make use of prediction algorithm on these pattern data to classify that weather a particular transaction is fraudulent or genuine.

When credit card is being used by unauthorized user the neural network based fraud detection system check for the pattern used by the fraudster and matches with the pattern of the original card holder on which the neural network has been trained, if the pattern matches the neural network declare the transaction ok

When a transaction arrives for authorization, it is characterized by a stream of authorization data fields that carry information identifying the cardholder (account number) and characteristics of the transaction (e.g., amount, merchant code). There are additional data fields that can be taken in a feed from the authorization system (e.g., time of day). In most cases, banks do not archive logs of their authorization files. Only transactions that are forwarded by the merchant for settlement are archived by the bank's credit card processing system. Thus, a data set of transactions was composed from an extract of data stored in Bank's settlement file. In this extract, only that authorization information that was archived to the settlement file was available for model development.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Fraud Detection

Matching the pattern does not mean that the transaction should exactly match with the pattern rather the neural network see to what extent there exist difference if the transaction is near by the pattern then the transaction is ok otherwise if there is a big difference then the chance of being a transaction illegal increase and the neural network declare the transaction a fault transaction.

The neural network is design to produce output in real value between 0 and 1 .If the neural network produce output that is below .6 or .7 then the transaction is ok and if the output is above .7 then the chance of being a transaction illegal increase.

There are some occasions when the transaction made by a legal user is of a quite different and there are also possibilities that the illegal person made use of card that fit into the pattern for what the neural network is trained. Although it is rare, yet If the legal user can't complete a transaction due to these limitation then it is not much about to worry But what about the illegal person who is making use of card , hare also work human tendency to some extent when a illegal person gets a credit card he is not going to make use of this card again and again by making number of small transaction rather he will try to made as large purchase as possible and as quickly that may totally mismatch with the pattern for what the neural network is trained.

TRANSACTION FRAUD SCORER

The neural network used in this fraud detection a three-layer, feed-forward network that use two training passes through the data set. The fast training pass involves a process of prototype cell commitment in which exemplars from the training set are stored in the weights between the first and second (middle) layer cells of the network. A final training pass determines local a posteriori probabilities associated with each of these prototype cells. P-RCE training is not subject to problems of convergence that can afflict gradient-descent training algorithms. The P-RCE network and networks like it have been applied to a variety of pattern recognition problems both within and beyond the field of financial services, from character recognition to mortgage underwriting and risk assessment layer consisted of a single cell that outputs a numeric response that can be considered as a "fraud score". This is analogous to credit scoring systems that produce a score, as opposed to a strict probability. The objective of the neural network training process is to arrive at a trained network that produces a fraud score that gives the best ranking of the credit card transactions. If the ranking were perfect, all of the high scoring transactions down to some threshold would be fraud; below this threshold, only good transactions would be ranked.

However, perfect separation of frauds from goods is not possible due to the inherently non-separable nature of the fraud and good distributions in the selected pattern recognition Space.

Final evaluation of the trained network can be done on the Blind Test data set. The Blind Test data represented an unsampled set of all Banks' transactions during last few months.

Learning Algorithm (Feed Forward Back Propagation)

The back propagation learning rule is a standard learning technique. It performs a gradient descent in the error/ weights space. To improve the efficiency, a momentum term is introduced, which moves the correction of the weights in the direction compliant with the last weight correction?

It is a multi-layer feed forward network that is trained by supervised learning.

Supervised learning means that the network is repeatedly presented with input/output pairs (I,O) provided by a supervisor, where O is the output the network should produce when presented with input I. These input/output pairs specify the activation patterns of the input and output layer. The network has to find an internal representation that result in the wanted input/output behavior. To achieve this, back propagation uses a two-phase propagates-adapt cycle.

i. First Phase: In the first phase the input is presented to the network and the activation of each of the nodes (processing elements) of the input layer is propagated to the hidden layer, where each node sums its input and propagates its calculated output to the next layer. The nodes in the output layer calculate their activations in the same way as the nodes in the hidden layer.

ii. Second Phase: In the second phase, the output of the network is compared with the desired output given by the supervisor and for each output node the error is calculated. Then the error signals are transmitted to the hidden layer where for each node its contribution to the total error is calculated. Based on the error signals received, connection weights are then *adapted* by each node to cause the network to converge toward a state that allows all the training patterns (input/output pairs) to be encoded.

PROBLEM WITH THE TRAINING OF NEURAL NETWORK

Problem with neural networks is that a number of parameter has to be set before any training can begin. However, there are no clear rules how to set these parameters. Yet these parameters determine the success of the training. In the most general case, neural networks consist of an (often very high) number of neurons, each of which has a number of inputs which are mapped via a relatively simple function to its output.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Networks differ in the way their neurons are interconnected (topology), in the way the output of a neuron determined out of its inputs (propagation function) and in their temporal behavior (synchronous, asynchronous or continuous).

The topology of a network has a large influence on the performance of that network but, so far, no method exists to determine the optimal topology for a given problem because of the high complexity of large networks. the choice of the basic parameter (network topology, learning rate, initial weights) often already determines the success of the training process. The selection of these parameters follow in practical use rules of thumb, but their value is at most arguable.

Genetic Algorithms Overview

The biological metaphor for genetic algorithms is the evolution of the species by survival of the fittest, as described by Charles Darwin. In a population of animals or plants, a new individual is generated by the crossover of the genetic information of two parents.

The genetic information for the construction of the individual is stored in the DNA. The human DNA genome consists of 46 chromosomes, which are strings of four different bases, abbreviated A, T, G and C. A triple of bases is translated into one of 20 amino acids or a "start protein building" or "stop protein building" signal. In total, there are about three billion nucleotides. These can be structured in genes, which carry one or more pieces information about the construction of the individual. However, it is estimated that only 3% of the genes carry meaningful information, the vast majority of genes - the "junk" genes - is not used.

The genetic information itself, the genome, is called the *genotype* of the individual. The result, the individual, is called *phenotype***.** The same genotype may result in different phenotypes. Twins illustrate this quite well.

Genetic algorithms are algorithms for optimization and machine learning based loosely on several features of biological evolution. They require five components:

i. A way of encoding solutions to the problem on chromosomes.

ii. An evaluation function which returns a rating for each chromosome given to it

Iii.A way of initializing the population of chromosomes.

iv. Operators that may be applied to parents when they reproduce to alter their genetic composition. Standard operators are mutation and crossover Parameter settings for the algorithm, the operators, and so forth.

GENETIC ALGORITHM ALONG WITH NEURAL NETWORK

(GANN) By combining genetic algorithms with neural networks (GANN), the genetic algorithm is used to find these parameters. The inspiration for this idea comes from nature:

In real life, the success of an individual is not only determined by his knowledge and skills, which he gained through experience (the neural network training), it also depends on his genetic heritage (set by the genetic algorithm). One might say, GANN applies a natural algorithm that proved to be very successful on this planet: It created human intelligence from scratch. The main question is how exactly GA and NN can be combined, i.e. especially how the neural network should be represented to get good results from the genetic algorithm

Information about the neural network is encoded in the genome of the genetic algorithm. At the beginning, a number of random individuals are generated. The parameter strings have to be evaluated, which means a neural network has to be designed according to the genome information. Its performance can be determined after training with back-propagation. Some GANN strategies rely only on the GA to find an optimal network; in these, no training. take place. Then, they are evaluated and ranked. The fitness evaluation may take more into consideration than only the performance of the individual.

Principle Structure of GA and GANN System

Individual's version and the network pruning algorithm. The first uses just the weight-encoding bits, the second merely the index-bit. For the later, the weight values of an already generated optimal network are used, the goal is to find a minimal network with good performance. Of course, the number of weights pruned has to be considered in the fitness function. GENITOR requires that a basic (maximal) architecture has to be designed for each problem. The resulting encoding format is a bit-string of fixed length.

The standard GA has no difficulties to deal with this genome. Since crossover can take place at any place of the bit string, a child may have a different weight value than either one of the parents. So, topology and weight values are optimized at the same time. Whitley reports that GENITOR tends to converge to a single solution, the diversity is reduced fast. It seems to be a good "genetic hill-climber". The approach was applied to simple Boolean functions.

CONCLUSION

In this paper we saw different technique that is being used to execute credit card fraud how credit card fraud impact on the financial institution as well as merchant and customer, fraud detection technique used by VISA and MasterCard. Neural network is a latest technique that is being used in different areas due to its powerful capabilities of learning and predicting.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this thesis we try to use this capability of neural network in the area of credit card fraud detection as we know that Back propagation Network is the most popular learning algorithm to train the neural network so in this paper BPN is used for training purpose and then in order to choose those parameter (weight, network type, number of layer, number of node e.t.c) that play an important role to perform neural network as accurately as possible, we use genetic algorithm, and using this combined Genetic Algorithm and Neural Network (GANN) we try to detect the credit card fraud successfully. The idea of combining Neural Network and genetic Algorithm come from the fact that if a person is inherently very talented and he is trained properly then chances of individual of success is very high.

REFERENCES

Abidi, S. S. R. (2001). Knowledge management in healthcare: towards 'knowledge-driven' decision-support services. *International Journal of Medical Informatics*, *63*(1-2), 5–18.

Ahmad, R., Kausar, A. R., & David, P. (Writer) (2007). *The social management of embodied knowledge in a knowledge community*.

Anderson, R. A., & McDaniel, R. R. (2000). Managing health care organizations: where profes-sionalism meets complexity science. *Health Care Management Review*, *25*(1), 83–92.

Andreas, R. (2005). Three-dozen knowledge-sharing barriers managers must consider. *Jour-nal of Knowledge Management*, *9*(3), 18–35.

Andreas, R., & Nicholas, L. (2006). Knowledge management in the public sector: stakeholder partnerships in the public policy development. *Journal of Knowledge Management*, *10*(3), 24.

Ansell, C. (2007). Fostering Innovation and Col-laboration. *Medical Device Technology*, *18*(1), 52.

Bali, R. K., & Dwivedi, A. N. (Eds.). (2007). *Healthcare Knowledge Management*: Springer.

Batalden, P., & Splaine, M. (2002). What will it take to lead the continual improvement and in-novation of health care in the twenty-first century? *Quality Management in Health Care*, *11*(1), 45.

Bate, S. P., & Robert, G. (2002). Knowledge management and communities of practice in the private sector: lessons for modernizing the National Health Service in England and Wales. *Public Administration*, *80*(4), 643–663.

Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., & Petersen, L. A. (1997). The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *Journal of the American Medical Association*, *277*(4), 307–311. doi:10.1001/jama.277.4.307

Bontis, N. (1996). There's a price on your head: Managing Intellectual Capital Strategically. *Business Quarterly*, (Summer): 40–47.

Bontis, N. (2001). Assessing knowledge assets: a review of models used to measure intellectual capital. *International Journal of Management Reviews*, *3*(1), 41–60.

Brian, D. P. (2006). *Final Report of the Federal Advisor on Wait Times*. Ottawa, Canada: Health Canada.

Buchan, I. E., & Hanka, R. (1997). Exchanging clinical knowledge via Internet. *International Journal of Medical Informatics*, *47*(1-2), 39–41.

Burnett, S. M., Williams, D. A., & Webster, L. (2005). Knowledge support for interdisciplin-ary models of healthcare delivery: a study of knowledge needs and roles in managed clinical networks. *Health Informatics Journal*, *11*(2), 146–160.

Caldwell, D. F., Chatman, J., O'Reilly, C. A., III, Ormiston, M., & Lapiz, M. (2008). Implementing strategic change in a health care system: the impor-tance of leadership and change readiness. *Health care management review*, *33*(2), 124(110).

Canadian Health Services Research Foundation. (2003). *The theory and practice of knowledge*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REINFORCEMENT LEARNING A TOOL FOR FILTERING PERSONALIZED WEB DOCUMENT

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem; and Inyiama, Hycient

[1] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[2] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

Document filtering is increasingly deployed in Web environments to reduce information over-load of users. We formulate online information filtering as a reinforcement learning problem, i.e. TD(0). The goal is to learn user profiles that best represent his information needs and thus maximize the expected value of user relevance feedback. A method is then presented that acquires reinforcement signals automatically by estimating user's implicit feedback from direct observations of browsing behaviors. This “learning by observation” approach is contrasted with conventional relevance feed-back methods which require explicit user feedbacks. Field tests have been performed which involved 10 users reading a total of 18,750 HTML documents during 45 days. Compared to the existing document filtering techniques, the proposed learning method showed superior performance in information quality and adaptation speed to user preferences in online filtering.

Keywords
Web based, Document, Reinforcement learning

1. INTRODUCTION

With the rapid progress of computer technology in recent years, electronic information has been explosively increased. This trend is especially remarkable on the Web. As the availability of the information increases, the need for finding more relevant information on the Web is growing [Belkin and Croft, 1996]. Currently, there are two major ways of accessing information on the Web. One is to use Web index services such as AltaVista, Yahoo, and Excite. The other is to manually follow or browse the hyperlinks of the documents by a user himself. However, these methods have some drawbacks. Since Web-index services are based on general purpose indexing methods, much of the retrieval results may be irrelevant to user's interests. In addition, manual browsing involves much time and efforts. High-quality information services require to capture the personal interests of individual users during the interaction with the information retrieval systems.

Several methods have been proposed to reflect user preferences. A classical approach is the Rocchio method [Rocchio, 1971] and its variants. This is a batch algorithm that modifies the original query vector by the vectors of the relevant and irrelevant documents. However, the batch algorithms tend to put large demands on memory and are slow in adaptation, thus not well suited to on-line applications. Recently, several on-line learning algorithms have been used for information retrieval and filtering. These include the Widrow-Hoff rule [Lewis et al., 1996] and the exponentiated gradient algorithm [Callan, 1998]. These algorithms learn training examples one at a time and thus more appropriate for learning in online fashion. However, all these methods have a drawback that the user has to provide explicit relevance feedback for the system to learn. Since providing relevance feedbacks is a tedious process and users may be unwilling to provide them, the learning capability of the filtering systems may be severely limited.

In this paper, we present a personalized information filtering method that learns user's interests by observing his or her behaviors during the interaction with the system. First, the system is trained on the explicit feed-back from the user. After this learning phase, the system estimates the relevance feedback implicitly based on the observations of user actions. This information is used to modify the user profiles. We regard filtering as a goal-directed learning process based on interactions with the environment. The objective is to maximize the expected value of the cumulative relevance feedback it receives in the long run from the user. This process is formulated as TD(0) learning, a general form of reinforcement learning [Sutton and Barto, 1998]. In this formulation, filtering is viewed as an interactive process which involves a generate-and-test method whereby the agent try actions,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

observe the outcomes, and selectively retain those that are the most effective. The advantage of TD(0) over other reinforcement learning methods is that it can learn without excessive delay of rewards. This is an important property in real-time interactions with the user in Web browsing environments. Additional feature of our approach is that it is learning by experimentation, in contrast to learning by instruction as adopted in most supervised learning methods. The method was implemented as WAIR (Web Agents for Information Retrieval), a platform for Web-based personalized information filtering services [Seo and Zhang, 2000].

Personalized Filtering as Reinforcement Learning

Information Filtering in WAIR

WAIR (Web Agents for Information Retrieval) was originally designed as a platform for the development of personalized information services on the Web. WAIR consists of three agents: an interface agent, a re-trieval agent, and a filtering agent. The interaction between the agents is illustrated in Figure 1. The overall procedure is summarized in Figure 2.

Initially, the user provides the system with a profile (Step 1). Typically, the initial profile consists of a few keywords. Then, the retrieval agent constructs a query using the profile and get N URLs (Step 2). Existing Web search engines are used to obtain the relevant URLs. The documents for the URLs are then retrieved and preprocessed, and their relevance values are estimated. The N documents are ranked, and M of them are filtered and presented to the user (Step 3). To balance exploration and exploitation, WAIR chooses the highest-ranked documents most of the time, but occasionally (with probability) it filters lower-ranked documents.

The interface agent observes user behavior and measures user feedback (Step 4). Two different types of user feedbacks are distinguished in WAIR. One is the “explicit” feedback in the form of scalar values to evaluate the relevance of the documents. This is provided by the user during the initial learning phase. A second type of feedback is the “implicit” feedback. This is not provided by the user, but estimated by the interface agent in WAIR .

That is, the users read filtered HTML documents by performing normal browsing behaviors, such as scrolling thumb up and down, bookmarking an URL, following the hyperlinks in the filtered document, and the WAIR infers from the behaviors how much the user was interested in each filtered document with a multi-layer neural network. This process is described in detail in the next section. The feedback information is then used to update the user profile (Step 5). Basically, this consists of inserting new terms, removing existing terms, and adjusting term weights of profile terms using the terms in the

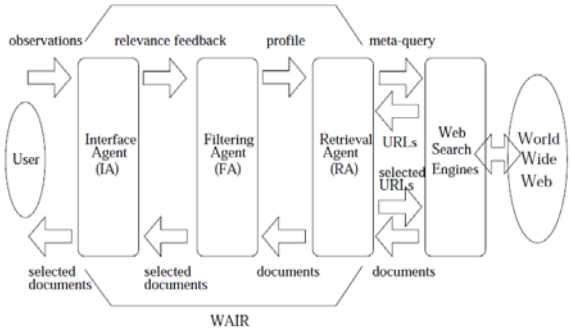


Figure 1: System architecture of WAIR

1. Get the initial profile from the user. Set $t \leftarrow 0$.
2. (Retrieval) Generate a query from the profile to retrieve N URLs.
3. (Filtering) Evaluate the relevance of documents.
   Rank the N documents and present M of them to the user.
4. (Interface) Get the feedback by observing user behavior.
5. (Learning) Update the user profile.
6. Set $t \leftarrow t + 1$. Go to step 2.

Figure 2: The overall procedure of WAIR

relevant/irrelevant documents. Then, the revised profile is used to get new documents by going to the retrieval step. Note that, the user provides only an initial query and then WAIR automatically retrieves and filters documents by observing user behaviors implicitly.

Filtering as Reinforcement Learning

The task of information filtering in WAIR is formulated as a reinforcement learning problem. Reinforcement learning is about learning from interaction how to behave in order to achieve a goal. The reinforcement learning agent and its environment interact over a sequence of discrete time steps. The actions are the choices made by the agent. The states are the basis for making the choices.

The rewards are the basis for evaluating choices. In WAIR, actions are defined as the decision-making as to whether to present the document to the user or not. States are defined as the pairs of the profile and the document to be filtered.

The policy is a stochastic rule by which the agent selects actions as a function of states. Formally, a policy is a mapping from each state

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

s and action as to the probability (s; a) of taking action a when in states. We use an -greedy policy for choosing an action given a state. That is, most of the time WAIR chooses the highest-ranked documents, but with probability, it chooses lower-ranked documents too. The rationale behind this policy is that it combines exploitation and exploration of search behavior. The selection of documents with the highest relevance value corresponds to exploitation of known information, while selecting random documents encourages exploration of unknown regions to find interesting documents which are unexpected by the user. An advantage of the -greedy method is that, in the limit as the number of actions increases, the probability of selecting the optimal action converges to greater than 1 , , i.e., to near certainty [Sutton and Barto, 1998].

The filtering agent's objective is to maximize the amount of reward it receives over time. The return is the function of future rewards that the agent seeks to maximize. Value functions of a policy assign to each state, or state-action pair, the expected return from that state, or state-action pair, the largest expected return achievable by any policy. The agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses action a_t to maximize the expected discounted return:

$$\begin{aligned} R_t &= r_{t+1} + \gamma r_{t+2} + +\gamma^2 r_{t+3} \cdots \\ &= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \end{aligned}$$

where γ is a parameter, $0 \leq \gamma \leq 1$, called the discount rate.

To make decisions on whether or not filter the documents, it is necessary to estimate value functions, i.e., functions of states that estimate how good it is to be in a given state. The notion of how good here is defined in terms of future rewards that can be expected, i.e. in terms of expected return. Value functions are defined with respect to particular policies. Informally, the value of a states under a policy , denoted V (s), is the expected return when starting in s and following thereafter. We can define V (s) as

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\{R_t \mid s_t = s\} \\ &= E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s \right\} \\ &= E_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\middle|\, s_t = s \right\} \\ &= E_{\pi}\{ r_{t+1} + \gamma V^{\pi}(s_{t+1}) \mid s_t = s \}, \end{aligned}$$

Learning Profiles from Implicit Feedbacks

In this section, we first describe the retrieval of documents in WAIR. Then, the procedures for estimating user feedbacks and updating user profiles are described.

Document Retrieval

The task of the retrieval agent is to get a collection of candidate HTML documents to be filtered. The retrieved documents undergo preprocessing. We use standard term-indexing techniques, such as removing stop-words and stemming [Frakes and Baeza-Yates, 1992]. Formally, a document is represented as a term vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{i,k}, ..., x_{i,d}),$$

where $x_{i;k}$ is the numeric value that term k takes on for document i, d is the number of terms used for document representation. In this work, we assume that $x_{i;k}$ represents the normalized term frequency, i.e. $x_{i;k}$ is proportional to the number of term k appearing in document i and k x_i k= 1. This is contrasted with the usual *tf idf* (term frequency inverse document frequency) [Salton, 1989] based indexing method in conventional information retrieval. We use only tf information because we focus on information filtering from a stream of Web documents. In contrast to the conventional information retrieval environments where the collection of documents is static over a long period of time, our situation addresses a dynamically changing environment. In this dynamic environment, the inverse document frequency (which is computed with respect to a static collection of documents) is not significant.

The ultimate goal of WAIR is to filter documents that best reflect user's preferences. This is done by learn-ing the profiles of users. A user profile consists of one or more topics. Topics represent user's information needs. In this section, we assume for simplicity that a profile consists of a single topic. The method can readily be generalized to multiple topics for a user by maintaining multiple profiles. Formally, the profile p is represented as a weight vector

$$\mathbf{w}_p = (w_{p,1}, w_{p,2}, ..., w_{p,k}, ..., w_{p,d}),$$

where $w_{p;k}$ is the weight of the kth term in the profile and k w_p k= 1. d is the number of terms used for describing the profiles. Formally, it is the same as the number of terms for representing documents. In WAIR, however, the maximum number of non-zero terms in the profile is limited to m < d. This is useful for concise description of user interests. Initially, the profile w_p contains only

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

a small number of non-zero terms that are contained in the original user query. The subsequent retrieval and user-feedback process expands and updates the number and weights of the profile terms, as described below.

WAIR searches the Web-documents by using existing Web-index services, i.e. AltaVista, Excite, and Lycos. That is, it formulates a query q_p that is forwarded to one or more Web search engines. Queries are constructed by choosing terms from the profile based on an -greedy selection method. The retrieval agent then selects *N* URLs from different engines and ranks them. The rank of document i for profile p is based on its similarity (or relevance) to the profile and computed as the inner product:

$$V(s_i) = \mathbf{w}_p \cdot \mathbf{x}_i = \sum_{k=1}^{d} w_{p,k} x_{i,k},$$

where $w_{p;k}$ and $x_{i;k}$ are the kth terms in profile p and document, respectively. The candidate documents are then sorted in descending order of $V_i(s_i)$, and M of them are presented to the user. Note that since the term vectors are normalized to $w_p = 1$ and $x_i = 1$, the relevance value is equivalent to the cosine correlation, i.e.

$$V(s_i) = \frac{\mathbf{w}_p \cdot \mathbf{x}_i}{\| \mathbf{w}_p \| \| \mathbf{x}_i \|}$$

$$\text{where } \| \mathbf{x}_i \| = \sqrt{\sum_{k=1}^{d} x_{i,k}^2}.$$

Conclusions

In this paper, we formulated the problem of information filtering as a TD(0) reinforcement learning problem, and presented a personalized Web-document filtering system that learns to follow user preferences from observations of his behaviors on the presented documents. A practical method was described that estimates the user's relevance feedback from user behaviors such as reading time, bookmarking, scrolling, and link-following actions.

Our experimental evidence from a field test on a group of users supports that the proposed method effectively adapts to the user's specific interests. This confirms that “learning from shoulders of the user” through self-generated reinforcement signals can significantly improve the performance of information filtering systems. In a series of short-term filtering environments, WAIR achieved superior performance when compared to the conventional feedback methods, including Rocchio, WH, and EG. In terms of adaptation speed, the proposed method converged to the user's specific interest faster than existing relevance feedback methods.

Our work has focused on personalizing information filtering based on existing Web-index services, i.e. AltaVista, Excite, and Lycos. Through the use of learning-based personalization techniques, WAIR could improve the quality of information service of the existing Web search engines. Since every search engine has its strengths and weaknesses, the meta-search approach of WAIR combines the strengths of different search engines while reducing their weaknesses. For the convenience of implementation, we used the conventional search engines directly. Using meta-search engines would further increase the final performance. Similar idea can be used to improve the quality of other Web information service systems.

The online nature of reinforcement learning makes it possible to approximate optimal action policies in ways that put more effort into learning to make good decisions for frequently encountered states, at the expense of less effort for infrequently encountered states.

This is the key property that distinguishes reinforcement learning from other relevance feedback methods based on supervised learning. Our experimental result confirms this view: information filtering is dictated by online adaptation based on a small number of documents. The reinforcement learning formulation gave more emphasis on decision making as to filtering the documents rather than just to learn the mappings or profiles. This resulted in better performance than simple supervised learning methods in the dynamic environments. Our work suggests that reinforcement learning can provides a better framework for personalization of information service in the Web environments than conventional supervised learning formulation.

In spite of our success in learning the user preferences in the WAIR system, it should be mentioned that the success comes in part from the environments where we made our experiments.

One is that the topics used for experiments were usually scientific and thus the filtered documents contained relatively less-ambiguous terms than those that might be contained in other usual Web documents. Another reason might be that the duration of our experiments were not very long during which the user interests did not change very much. The adaptation to user's interests during a longer period of time in a more dynamic environment should still be tested. From a more practical point of view, the response time is a crucial factor in the information retrieval and filtering.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

However, our focus in this paper was confined to the relevance feedback. Learning from users to minimize their response time is one of our research topics in the future.

REFERENCES:

Belkin, N.J. and Croft, W.B. 1992. Information filtering and information retrieval: Two sides of the same coin, *Communications of the ACM*, 35(12):29-38.

Boyan J., Freitag D., and Joachims T. 1996. A machine learning architecture for opti-mizing Web search engine, In *Proc. AAAI Workshop on Internet-Based Information Systems*, pp. 324-335.

Callan J. 1998. Learning while filtering documents, In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR-98), pp. 224-231.

Falk, A. and Josson, I.M. 1996. PAWS: An agent for WWW-retrieval and filtering, In *Proc. Practical Application of Intelligent Agents and Multi-agents Technology* (PAAM-96), pp. 169-179.

Frakes, W.B. and Baeza-Yates, R. 1992. Stemming algorithms, In *Informa-tion Retrieval: Data Structures and Algorithms*, pp. 131-160, Prentice Hall.

Hirashima, T., Matsuda, N., Nomoto, T., and Toyoda, J. 1998. Context-sensitive filtering for browsing in hypertext, In *Proc. Int. Conf. on Intelligent User Interfaces* (IUI-98), pp. 119-126.

Joachims, T., Freitag D., and Mitchell, T.M. 1997. WebWatcher: A tour guide for the World Wide Web, In *Proc. Int. Joint Conf. on Artificial Intelligence* (IJCAI-97), pp. 770-777.

Kamba, T., Sakagami, H., and Koseki, Y. 1997. ANATAGONOMY: A personalized newspaper on the World Wide Web, *Int. Jor. of Human-Computer Studies*, Vol. 46, pp. 789-803.

Kindo, T., Yoshida, H., Morimoto, T., and Watanabe, T. 1997. Adaptive personal in-formation filtering system that organizes personal profiles automatically, In *the Proc. Int. Joint Conf. on Artificial Intelligence* (IJCAI-97), pp. 716-721.

Lashkari, Y., Metral, M., and Maes, P., 1994. Collaborative interface agents, In *Proc.of the Twelfth National Conf. on Artificial Intelligence* , pp.444-450.

Lewis, D.D., Schapire, R.E., Callan, J.P., and Papka, R. 1996. Training algorithms for linear text classifiers, In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR-96), pp. 298-306.

Lieberman, H. 1995. Letizia: An agent that assists Web browsing, In *Proc. Int. Joint Conf. on Artificial Intelligence* (IJCAI-95), pp. 475-480.

Maes, P. 1994. Agents that reduce work and information overload, *Communications of the ACM*, 37(7):31-40.

Mitchell, T.M. 1997. *Machine Learning*, McGraw-Hill.

Morita, M. and Shinoda, Y. 1994. Information filtering based on user behav-ior analysis and best match text retrieval, In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR-94), pp. 272-281.

Pazzani, M. and Billsus, D. 1997. Learning and revising user profiles: the iden-tification of interesting Web sites, *Machine Learning*, 27:313-331.

Rocchio, J.J. 1971. Relevance feedback in information retrieval, In *The SMART Retrieval System*, Prentice Hall, pp. 313-323.

Salton, G. 1989. *Automatic Text Processing*, Addison Wesley.

Salton, G. and Buckley, C. 1990. Improving retrieval performance by relevance feedback, *Journal of American Society for Information Science*, 41:288-297.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Influence of Distributed Generation on the Protection Scheme of Local Distribution System

Usman Yousaf[1], Dr. Azzam ul Asar[2], Alishpa Khattak[3]
Department of Electrical Engineering, CECOS University, Peshawar, Pakistan
emails: usm_usf@hotmail.com[1], dean@cecos.edu.pk[2], alishpakhattak92@gmail.com[3]

***Abstract*— The integration of Distributed Generation (DG) into the distribution system is considered as an achievement made in the field of power system. With such penetration of DG requires assessment of protection schemes used in traditional distribution system. Short circuit studies are needed to be performed for determining an adequate protection scheme for such an integrated distribution system. This paper presents short circuit analysis of 132 KV Dargai Grid Station (GS) Pakistan incompliance to IEC 60909.The Distribution system is modeled in Electrical Transients and Analysis Program (ETAP) software and comparative fault analysis has been performed with and without DG. The fault location is made fixed while DG location is varied. It is found that there is significant increase in fault current with the DG Penetration and the fault current depends upon total feeder length, distance of fault location from the DG and Grid and extent of flowing current.**

***Index Terms*—Distributed Generation, ETAP, IEC 60909, Short Circuit Analysis, Distribution System**

I. INTRODUCTION

The yearly growing electrical energy demand has increased the penetration of DG significantly in to the distribution network. Distribution system is the link between the end user and the utility system [1]. Various benefits are provided to the utility and the consumer by interconnecting DG to an existing distribution system. DG provides an enhanced power quality, higher reliability of the distribution system and peak shaving. However, power system protection being one of the major issue several technical problems are associated with the integration of DG into existing distribution system. The radial power flow is lost and the fault level of the system is increased due to the incorporation of DG [2].

Short circuit studies are one of the most important tasks in power system analysis. According to IEC 60909 short circuit is the accidental or intentional conductive path between two or More conductive parts forcing the electric potential difference between these conductive parts become zero [3].

Short circuit currents produce powerful magnetic forces and intense heat in the power system, which can result in considerable damage to the power system protective equipment. As the breaking capacity of circuit breakers is described by the initial symmetrical fault current flows through the system .when fault occurs, these values of short-circuit currents must be determined to ensure that the short-circuit ratings of all equipment are adequate to sustain the currents available at their locations [3].

This paper aims to verify the effect of DG on the fault current contribution and also investigates viable location for DG coupling by considering three scenarios based on peak load, feeder length and fault location from DG of the distribution system using ETAP software [4], [5].

II. IEC 60909 SHORT CIRCUIT ANALYSIS

IEC 60909 Short Circuit Currents in Three Phase System describes an internationally accepted method for the calculation of fault currents. In applying the standard, two levels of fault based on voltage factor are typically calculated [6],[7].

- The maximum current which causes the maximum thermal and electromagnetic (Mechanical) effects on equipment and is used to determine the equipment rating.
- The minimum current which is used for the setting of protective devices such as relay settings and coordinated relay operation.

Depending on the position within the cycle at which the fault forms, a dc offset will be present, decaying overtime to zero. This creates an initial symmetrical short circuit I''_k, which will decay over time to the steady state short circuitI_k. [3]

A. Initial AC symmetrical short circuit fault current I''_k

The Maximum initial short circuit occurs for a system when three phase fault develop. This current is the root mean square value of the initial component of the short circuit current, which can be calculated by eq 1.[6]

$$I''_K = \frac{C.U_n}{\sqrt{3}\cdot Z_K} = \frac{C.U_n}{\sqrt{3}.\sqrt{R^2 \pm X^2}} \qquad (1)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Where
U_n=Nominal voltage
Z_K=equivalent short circuit impedance at the fault location
C=voltage correction factor

The "C" factor or voltage factor is the ratio of equivalent voltage to the nominal voltage, and required to account for variation due to time and place, transformer taps, static load and capacitance, generator and motor sub transient behavior. [3]

B. *Peak Short Circuit current I_P*

It is the maximum momentary value of the short circuit current. It is only calculated for maximum short circuit current and can be calculated by eq.2

$$\boldsymbol{I_P=\sqrt{2}\times k\times I_K''} \qquad (2)$$

"K" is the function of system R/X ratio at the fault location and can be calculated by eq.3

$$\mathbf{k=1.02+0.98.}e^{\frac{-3R}{X}} \qquad (3)$$

At a fault location the F, the total amount of peak short circuit current is the absolute value of all partial short circuit currents as shown by eq.4. When the R/X ratio remains less than 0.3 at all branches, the R/X ratio of equivalent impedance at the fault location can be used for calculation of k [6].

$$\boldsymbol{i_p=\sum_i i_{pi}} \qquad (4)$$

The system R/X ratio depends on the method selected for calculation. Method A is for uniform
X/R ratio. Method B is for meshed networks and Method C is for non-meshed networks. [3]

C. *Steady State Short Circuit Current I_K*

It is the value of short circuit current when several cycles have been passed. For calculation of maximum steady short circuit current the synchronous generator excitation is kept at maximum.

III. System Description And Simulation Model

For determining actual performance of the power system, the proper mathematical model, accurate parameters of the power network, the generators, transformers and actual loads have to be identified [3]. Actual data related to the transformers, generators, load, and electrical parameters is collected from the power houses and the grid. Fig. 1 shows single line diagram of 132 kV GS Dargai Pakistan simulated in ETAP software [8], [9].

The Dargai Grid station is connected to 81 MW Malakand III Hydro Power Complex which has three 27.2 MW Generator Units transmitting power by Two 132KV outgoing transmission lines. The Grid is also connected by single 132 KV incoming Transmission line to 20 MW Dargai Power House which consists of two 10 MW Generator Units. In Malakand III Hydro Power Complex three 32 MVA Power Transformers step up 11 kV Generated voltage to 132 KV and in Dragai power House two 15 MVA Power Transformers steps up 11 KV generated Voltage to 132 KV, Two 20/26 MVA Distribution Transformers are installed at 132 KV Dargai Grid station which steps down the incoming 132 KV into 11 KV.

Generating units in study are represented by detailed model, with transient and sub-transient circuits on both the direct and quadrate axes been considered, as it describes all possible contribution to the short circuit current.

The Grid station is also connected by two Transmission lines to Chakdara and Mardan Power grids having 568.18 MVA_{SC} and 793.65 MVA_{SC} capacity respectively.

Twelve 11 KV local radial distribution feeders are emanating from the GS to the consumers. The total real time maximum current on the Grid station is 3420 A.

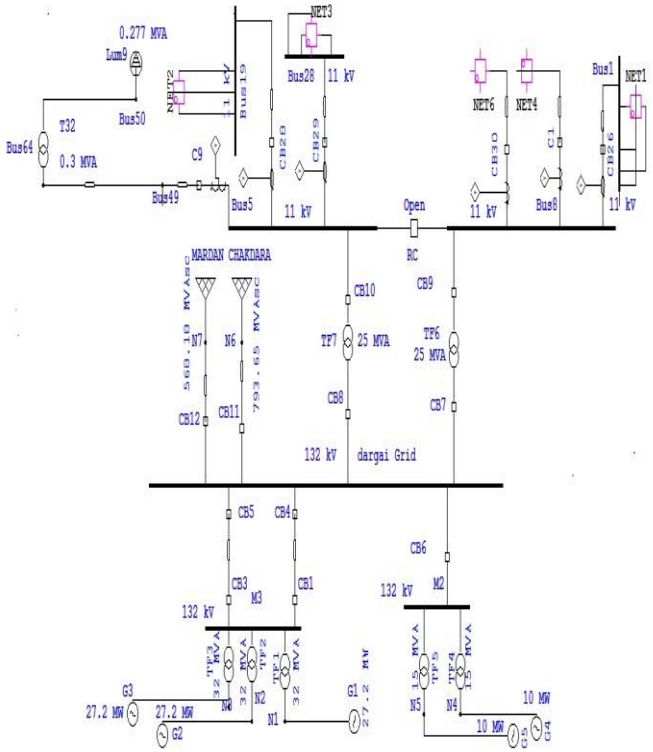


Fig.1 modeled single line diagram of 132 KV grid station Dargai with DG

IV. Case Analysis And Simulation Results

In this paper short circuit analysis is carried out on four feeders, firstly fault currents are obtained without any DG penetration while in other four cases different locations of DG are considered based on various feeder parameters so as to investigate the influence of DG in contribution of short circuit current under fault condition, the case wise simulation results are described below [10].

A. *Fault Analysis without DG.*

Case 1:

The system is simulated for a fixed fault without any DG penetration. The table of results for this case is listed in table I. Nominal voltage =11kv voltage factor c=1.1(max) fault current is in Kilo Ampere

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I
SHORT CIRCUIT REPORT WITHOUT DG

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	1.934	1.944	1.944
Peak Current (KA,rms)	3.6	3.325	3.579	3.345
Breaking Current(KA,rms,Sym.)	1.96	1.934	1.944	1.944
Steady State Current (KA,rms)	1.96	1.934	1.944	1.944

B. Fault Analysis with DG

Wind turbine Generators having equal ratings of 2MW are considered as DG source for penetration at different nodes in the selected feeders. The four cases under studied are described as under.

Case 2: 2 MW DG located at Bus 10
The system is simulated for a fixed fault with 2MW DG source connected to a 400A load feeder bus comprising of 70Km length. The table of results for this case is listed in table II.

TABLE II
SHORT CIRCUIT REPORT WITH DG AT BUS 10

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	3.06	1.934	1.955	1.945
Peak Current (KA,rms)	5.91	3.326	3.588	3.345
Breaking Current(KA,rms,Sym.)	3.06	1.934	1.955	1.945
Steady State Current (KA,rms)	3.06	1.934	1.955	1.945

Case 3: 2 MW DG located at bus 23
The system is simulated for a fixed fault with 2MW DG source connected to 80A load feeder bus comprising of 50KM length. The table of results for this case is listed in table III.

TABLE III
SHORT CIRCUIT REPORT WITH DG AT BUS 23

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	3.222	1.944	1.956
Peak Current (KA,rms)	3.6	5.98	3.579	3.377
Breaking Current(KA,rms,Sym.)	1.96	3.222	1.944	1.956
Steady State Current (KA,rms)	1.96	3.222	1.944	1.956

Case 4: 2MW DG located at Bus 38
The system is simulated for a fixed fault with 2MW DG source connected to 80A, 70Km feeder. The result for this case is listed in table IV.

TABLE IV
SHORT CIRCUIT REPORT WITH DG AT BUS 38

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	1.934	2.099	1.945
Peak Current (KA,rms)	3.6	3.326	3.996	3.345
Breaking Current(KA,rms,Sym.)	1.96	1.934	2.099	1.945
Steady State Current (KA,rms)	1.96	1.934	2.099	1.945

Case 5:2 MW DG located at Bus 49
In this case 2MW DG source is connected to a feeder having 50km length and possessing 400A current. The results are listed in table 5

TABLE V
SHORT CIRCUIT REPORT WITH DG AT BUS 49

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	1.947	1.944	3.722
Peak Current (KA,rms)	3.6	3.36	3.579	6.998
Breaking Current(KA,rms,Sym.)	1.96	1.947	1.944	3.722
Steady State Current (KA,rms)	1.96	1.947	1.944	3.722

V. Comparative Analysis and Discussion

This section describes the Comparison of initial symmetrical fault current, peak short circuit current and steady state current of the system during fault with and without the DG interconnection firstly while in the second section the comparison of fault currents between these feeders is discussed briefly.

A. Comparison of fault current with and without DG connected

The below chart shows the values of short circuit currents at the studied buses and is considered as base and set values for the protection settings of equipment used in the grid, those values will be compared with the values obtained from all other cases which contains a DG source.

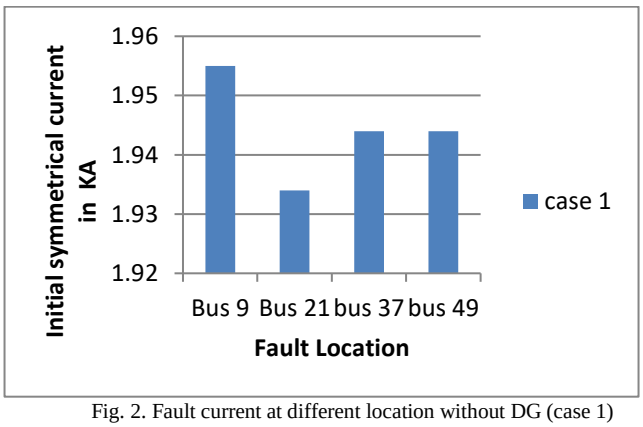


Fig. 2. Fault current at different location without DG (case 1)

To investigate the effect of DG on a feeder during fault conditions a series of four cases are considered which are compared with case 1; fig. 3 shows the comparison of case 2 with case 1.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

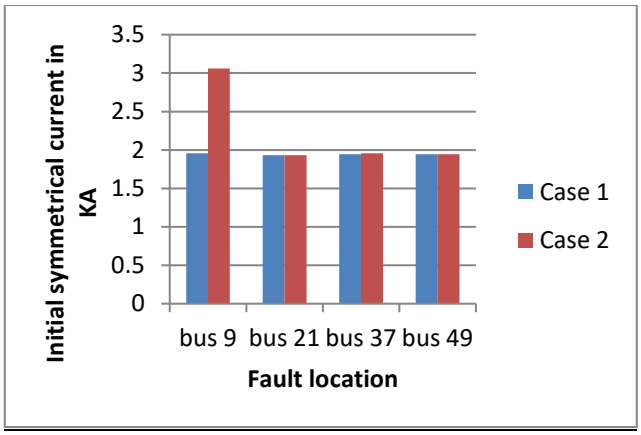


Fig. 3. Comparison between case 1 and 2

It shows that whenever DG is integrated at node 10 and a bolted fault occurs at bus 9 of the same feeder the magnitude of short circuit current increases from 1.955 KA to 3.061KA

In this case the parameters of feeder are different from the previous scenario, when a 2MW DG is brought into the system at node 23 and fault occurs at the bus 21 of the same feeder having length of 50KM and possessing 80A current, the results are compared with the base case as described in the fig. 4 below.

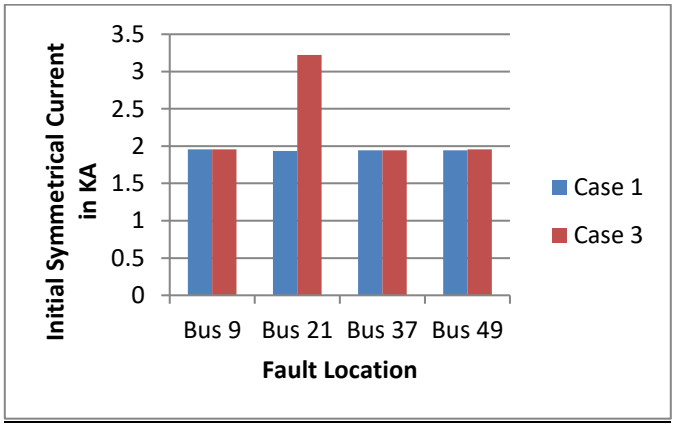


Fig. 4. Comparison between case 1 and 3

It is also evident from the comparison that addition of DG causes the short circuit contribution of feeder during fault conditions increased by collective 1.286KA.
The below comparison is validation of the objective that magnitude of short circuit current is increased during fault conditions with DG even when the length of feeder is maximum and minimum current of 80A is passing through it as shown in the fig. 5

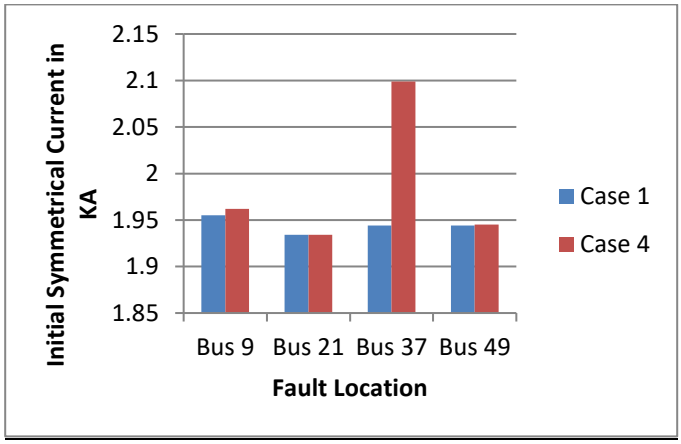


Fig. 5. Comparison between case 1 and 4

In the Last case DG is added at node 49 of a feeder comprising of 400A current and a total length of 70KM, the results in case 5 are compared with case 1 in the fig. 6 below.

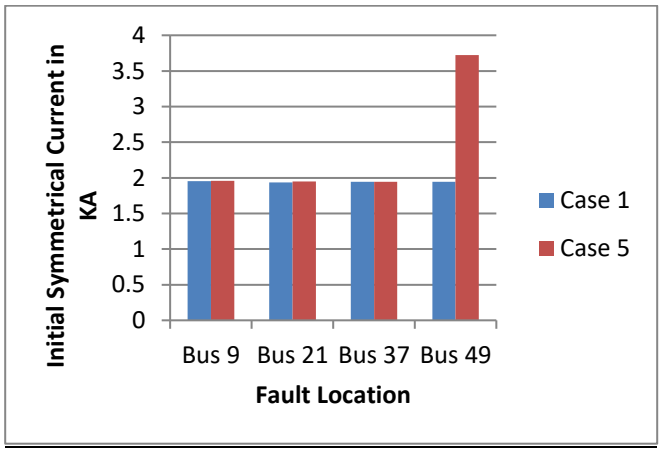


Fig. 6. Comparison between case 1 and 5

From the comparison it is extracted that whenever the location of fault and integrated DG is same the feeder contributes to maximum amount short circuit current during fault conditions, thus establishing a conclusion that whatever the parameter of a feeder may be an integrated DG will result in increase of short circuit current.

B. Comparison of Feeders containing DG

In this section two cases are considered for investigation to study the impact of DG on a feeder when its length is increased during fault conditions, the location of fault is maintained constant and short circuit currents are compared with each other as obtained in case 2 and case 5 which is described in the fig. 7

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

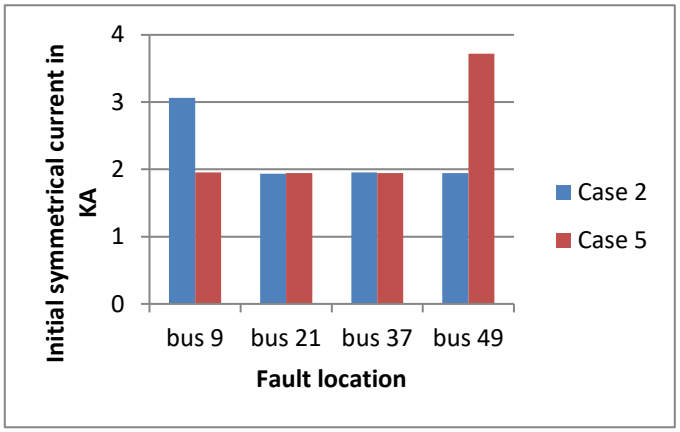


Fig. 7. Comparison between case 2 and 5

From the analysis it is obvious that when the length of a feeder is enhanced form 50km to 70Km while keeping maximum amperes of 400A flowing through it and fault occurs on the feeders integrated with DG, the short circuit current decrease from 3.722A to 3.061KA, so a decrease of 0.661KA suggests with conclusion that whenever the length of a line is increased irrespective of maximum or minimum load connected to it with DG integrated during fault, the fault current minimizes due to the fact that line impedance become larger with enhanced length thereby decreasing the effect of fault current on the distribution grid.

VI. Conclusion

Penetration of DG into a distribution system causes an increase in the fault level of the network at any fault location. In this paper Practical 132KV Grid Station has been considered as a case study for short circuit analysis with the DG connected. During the analysis it has been investigated that DG integration into the 11KV distribution feeder changes the initial symmetrical current contribution which can alter the protection configuration of the Grid station.

This paper also provides suitable location for DG integration in the selected feeders on the basis of its length suggesting that lengthy feeder has less short circuit current contribution and might be considered for penetration as it will not vary from the default protection setting of equipment's in the Grid.

VII. References

[1] Sarabia Angel Fernández. "Impact of distributed generation on distribution system." PhD dissertation Aalborg University, 2011.

[2] Jyotsna Sharma, Akhilesh Dobhal, " Analysis and Simulation of Faults in Power Distribution Systems with Distributed Generation *International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 9, September- 2014 ISSN: 2278-0181.*

[3] Renuka Kamdar, Mukesh Kumar Kirar, Manoj Kumar, Ganga Agnihotri, " Short Circuit Analysis of an Industrial Distribution System," *Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering* ISBN: 978-981-07-6260-5 doi:10.3850/ 978-981-07-6260-541.

[4] Hima A.Patel, Viabhav M.Sharma, and Anuradha Deshpande, " Relay Coordination using ETAP," *International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015 1583 ISSN 2229-5518.*

[5] Elmathana, Mohamed Talal Mohamed, Ahmed Zobaa, and George Smith. "The Effect of Distributed Generation on Power System Protection." (2010).

[6] Bong Eng Yueh, "Distributed Generation Impacts on the Coordination of Overcurrent Relay in Distributed Network," Universiti Teknologi Malaysia, Nov. 2012.

[7] "International standard IEC 60909, short circuit current in three phase a.c systems" International Electro technical commission, Jul. 2001.

[8] Khan, R. A. J., Junaid, M., & Asgher, M. M. (2009, November). Analyses and monitoring of 132 kV grid using ETAP software. In *Electrical and Electronics Engineering, 2009. ELECO 2009. International Conference on* (pp. I-113). IEEE.

[9] Keith Brown, Herminio Abcede, Farrokh Shokooh, "Interactive simulation of power system & ETAP application and Techniques" IEEE operation Technology, Irvine, California.

[10] Jillani, Syed Sagheer Hussain Shah. "Impact of distributed generation on distribution systems and its protection." (2012).

.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm

Abdeljalil EL ABDOULI, Larbi HASSOUNI, Houda ANOUN
RITM Laboratory, CED Engineering Sciences
Ecole Superieure de Technologie
Hassan II University of Casablanca, Morocco
elabdouli.abdeljalil@gmail.com, lhassouni@hotmail.com, houda.anoun@gmail.com

***Abstract*—Twitter is a web-based communication platform, which allows its subscribers to disseminate messages called "tweets" of up to 140 characters where they can share thoughts, post links or images. Therefore, Twitter is a rich source of data for opinion mining and sentiment analysis. The simplicity of use and the services offered by the Twitter platform allow it to be widely used in the Arab world and especially in Morocco, this popularity leads to an accumulation of a large amount of raw data that can contain a lot of valuable information. In this paper, we address the problem of sentiment analysis in Twitter platform. First, we try to classify the Moroccan users' tweets according to the sentiment expressed in them: positive or negative. Second, we discover the subjects related to each category to determine what they concern, and finally, we locate these "tweets" on Moroccan map according to their categories to know the areas where the tweets come from. To accomplish this, we adopt a new practical approach that applies sentiment analysis to Moroccan "tweets" using a combination of tools and methods which are: (1) Apache Hadoop framework (2) Natural Language Processing (NLP) techniques (3) Supervised Machine Learning algorithm "Naive Bayes" (4) Topic Modeling using LDA (5) Plotting tool for interactive maps called "Folium". The first task of our proposed approach is to automatically extract the tweets with emotion symbols (e.g., emoticons and emoji characters) because they directly express emotions regardless of used language, hence they have become a prevalent signal for sentiment analysis on multilingual tweets. Then, we store the extracted tweets according to their categories (positive or negative) in a distributed file system using HDFS (Hadoop Distributed File System) of Apache Hadoop framework. The second task is to preprocess these tweets and analyze them by using a distributed program written in Python language, using MapReduce of Hadoop framework, and Natural Language Processing (NLP) techniques. This preprocessing is fundamental to clean tweets from #hashtags, URLs, abbreviations, spelling mistakes, reduced syntactic structures, and many; it also allows us to deal with the diversity of Moroccan society, because users use a variety of languages and dialects, such as Standard Arabic, Moroccan Arabic called "Darija", Moroccan Amazigh dialect called "Tamazight", French, English and more. Afterward, we classify tweets obtained in the previous step using Naive Bayes algorithm into two categories (positive or negative), then we use the Topic Modeling algorithm LDA to discover general topics behind these classified tweets. Finally, we graphically plot classified tweets on our Moroccan map by using the coordinates extracted from them.**

Keywords: Apache Hadoop framework; HDFS; MapReduce; Python Language; Natural Language Processing; Supervised Machine Learning algorithm "Naive Bayes"; Topic Modeling algorithm LDA; Plotting tool for interactive maps.

I. INTRODUCTION

The emergence of Web 2.0 has led to an accumulation of valuable information and sentimental content in the Web; such content is often found in the comments of users of Social Network Platforms, in messages posted in discussion forums and product review sites, etc. The Twitter platform is very popular, and its users post a lot of comments to express their opinions, sentiments, and other information. This transforms twitter platform into a rich source of data for data mining and sentiment analysis. In this paper, we are interested in the sentiment analysis of the Moroccan users, we provide, below, some statistics on their activities. According to the Arab Social Media Report [1], which started in 2011 and aims to understand the impact of social media on societies, development, and governance in the Arab region, the monthly number of active users of the platform Twitter nearly doubled between 2014 and 2017. It went from 5.8 million to about 11.1 million. Regarding Morocco, the number of active users of the Twitter platform has grown of 146,300 users, in the last three years, to reach the number of 200 thousand users. Morocco thus ranks 9th among the Arab countries registering the highest number of users. These statistics prompted us to lead a study that aims to analyze the sentiments expressed in the tweets published by Moroccan users, despite the difficulties quoted before.

The primary aim of this research is to identify the sentiments contained in the tweets posted from the Moroccan region by proposing a new practical approach for analyzing the Moroccan user-generated data on Twitter. Our approach is based on a system, which automatically handles the streaming of the most recent tweets from Twitter platform using the open and accessible API of Twitter that returns well-structured tweets in JSON (JavaScript Object Notation) format. These tweets shape the training set, and are classified into two categories (Positive or Negative) according to the emotion symbols (e.g., emoticons and emoji characters) which exist in each tweet, then they are stored in our distributed system using HDFS [2]. These tweets are preprocessed by a distributed program using MapReduce [3], which is written in Python language using Natural Language Processing (NLP) techniques [4], and it's launched on MapReduce using the Pig UDF [5] (User Defined Functions). This preprocessing is fundamental to clean the tweets which are very noisy and contain all kind of spelling, grammatical errors and also to handle the linguistic diversity used by Moroccan users in the tweets. The result of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the previous step is a clean and filtered corpus of tweets that is divided into "Positive" (text with happy emoticons), and Negative" (text with sad and angry emoticons) samples. This corpus is used to form the training set for the Naive Bayes algorithm to identify the sentiment within the new collected tweets, then we apply topic modeling using LDA to discover the hidden topics within these tweets. Finally, we graphically plot the classified tweets using a tool called "Folium" on our Moroccan map by using the coordinates extracted from them, to discover the relationship between the areas of classified tweets and determined topics.

The remainder of the paper is organized as follows; we present some related work in Section II. In Section III, we introduce the tools and methods used to realize our system. In Section IV, we describe our system. Finally, in Section V; we end with a conclusion and work in perspective.

II. Related Work

Sentiment Analysis is receiving an increasingly growing interest from many researchers, which have begun to search various ways of automatically collecting training data and perform a sentiment analysis. [12] have relied on emoticons for defining their training data. [13] have used #hashtags for creating training data and they limit their experiments to sentiment/non-sentiment classification, rather than (positive-negative) classification. [14] have used emoticons such as ":-)" and ":-(" to form a training set for the sentiment classification, the author collected texts containing emoticons from Usenet newsgroups, and the dataset was divided into "positive" and "negative" samples. [15] have covered techniques and approaches that promise to enable the opinion-oriented information retrieval directly. [16] have used Twitter to collect training data and then to perform a sentiment search, they construct a corpus by using emoticons to obtain "positive" and "negative" samples and then use various classifiers, the best result was obtained by using Naïve Bayes Classifier.

III. Tools And Methods

A. Apache Hadoop

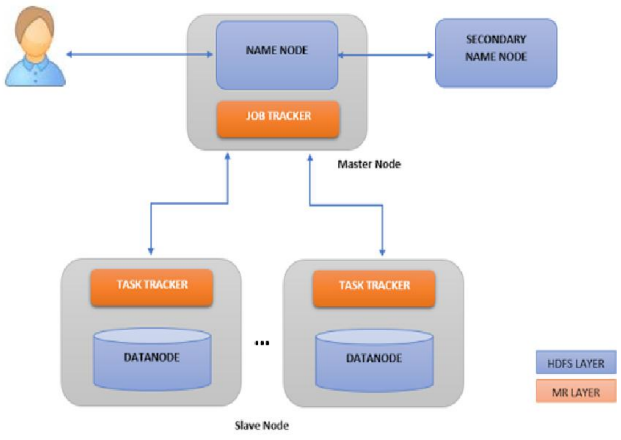


Figure 1. Apache Hadoop Architecture

Our approach is built using a specialized infrastructure, based on the Apache Hadoop Framework. The Apache Hadoop is an open-source software framework written in Java for processing, storing and analyzing large volumes of unstructured data on computer clusters built from commodity hardware.

The Hadoop Framework become a brand name, which contains two primary components. The first one is HDFS [5], which stands for Hadoop distributed file system; it is an open-source data storage, inspired by GFS (Google File System), it is a virtual file system that looks similar to any other file system, but the difference is that the file gets split into smaller files. The second one is MapReduce, which is an open-source programming model developed by Google Inc. Apache adopted the ideas of Google MapReduce and improved it. MapReduce provides a mechanism to break down every task into smaller tasks and the integration of results.

The HDFS (Hadoop Distributed File System) [2] system has many similarities with existing distributed file systems. However, the differences are significant, it is highly fault-tolerant and designed using low-cost hardware, also designed to be available and scalable. It provides high throughput access to stored data and can store massive files reaching the terabytes. By default, each stored file is divided into blocks of 64 MB, each block is replicated in three copies. The HDFS is based on Master and Slaves architecture in which the master is called the NameNode and slaves are called DataNodes, and it consists of:

a) Single NameNode: running as a daemon on the master node, it holds the metadata of HDFS by mapping data blocks to data nodes, and it is the responsible of managing the file system namespace operations.

b) Secondary NameNode: performs periodic checkpoints of the file system present in the NameNode and periodically joins the current NameNode image and the edits log files into a new image and uploads the new image back to the NameNode.

c) DataNodes: running as daemons on slave nodes, they manage the storing of blocks within the node (their default size is 128 MB). They perform all file system operations according to instructions received from the NameNode, and send a Heartbeat containing information about the total storage capacity of DataNode and Block report on every file and block they store to the NameNode.

The MapReduce [3] is the heart of Hadoop. It is a software framework that serves as the compute layer of Hadoop, it is modeled after Google's paper on MapReduce. It's characterized by fault tolerance, the simplicity of development, scalability, and automatic parallelization. It allows parallelizing the processing of massive stored data by decomposing the job submitted by the client into Map and Reduce tasks. The input of the Map task is a set of data as a key-value pair, and the output is another set of data as a key-value pair. The input of the reduce task is the output from a map task. Between the reduce input and the map output, MapReduce performs two essential operations, shuffle phase that covers the transformation of map outputs based on the output keys, and sort phase that includes the merge and sort of map outputs.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The MapReduce is also based on a master-slave architecture, and it consists of:

a) *JobTracker*: is running as a daemon on the master node, its primary role is accepting the job and assigning tasks to TaskTrackers running on slave nodes where the data is stored. If the TaskTracker fails to execute the task, the JobTracker assigns the task to another TaskTracker where the data are replicated.

b) *TaskTracker*: running as a daemon on slave nodes, it accepts tasks (Map, Reduce, and Shuffle) from JobTracker and executes program provided for processing. The TaskTrackers report the free slots within them to process data and also their status to the JobTracker by a heartbeat.

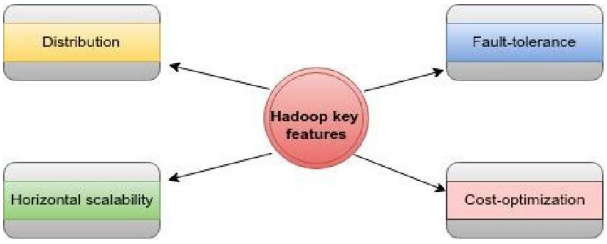


Figure 2. Key Features of Hadoop

The Key Features of Hadoop are:

Distribution: The storage and processing are spread across a cluster of smaller machines that work together.

Horizontal scalability: It is easy to extend a Hadoop cluster by adding new devices.

Fault-tolerance: Hadoop continues to operate even when a few hardware or software components fail to work correctly.

Cost-optimization: Hadoop runs on standard hardware; it does not require expensive servers.

Other Hadoop-related projects [7] at Apache that can be installed on top of or alongside Hadoop include:

- *Flume* [21]*:* is a framework for populating massive amounts of data into Hadoop.
- *Oozie* [22]*:* is a workflow processing system.
- *Mahout* [23]: Mahout is a data mining library.
- *Pig* [8]: a high-level data-flow language and execution framework for parallel computation.
- *Avro* [24]: a data serialization system.
- *HBase* [25]: a scalable and distributed database that supports structured data storage for large tables.
- *Hive* [26]: a data warehouse infrastructure that provides data summarization and ad hoc querying.
- *Spark* [27]*:* provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- And much more.

B. *Natural Language Processing (NLP)*

Natural Language Processing [4] is a part of computer science focused on developing systems that allow computers to recognize, understand, interpret and reproduce human language. NLP is considered as a subfield of artificial intelligence, and by using its algorithms, developers can perform tasks such as topic segmentation, translation, automatic summarization, named entity recognition, sentiment analysis, speech recognition, and much more.

There are two components of NLP. The first component is Natural Language Understanding (NLU) whose main function is to convert human language into representations that are easier for computer programs to manipulate. The other is Natural Language Generation (NLG) translate information from computer databases into readable human language. There are five steps in NLP:

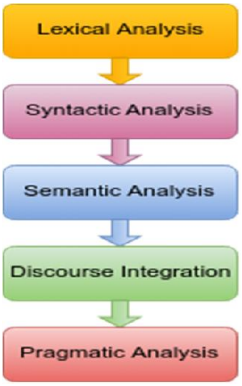


Figure 3. The steps of NLP

a) *Lexical Analysis:* identifying and analyzing the structure of words and dividing the whole text into paragraphs, sentences, and words.

b) *Syntactic Analysis:* analyzing and arranging words in a sentence in a structure that shows the relationship between them.

c) *Semantic Analysis:* extracting the exact meaning or the dictionary meaning of sentences from the text.

d) *Discourse Integration:* handles the meaning of current sentence depending on the sentence just before it.

e) *Pragmatic Analysis:* analyzing and extracting the meaning of the text in the context.

We use Natural Language Processing to perform tasks such as:

- Tokenization / segmentation
- Part of Speech (POS) Tagging: assign part-of-speech to each word.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Parsing: create a syntactic tree for a given sentence.
- Named entity recognition: recognize places, people...
- Translation: Translate a sentence into another language.
- Sentiment analysis.
- Etc.

Using the NLP is necessary for our system because tweets are characterized by a noisy text containing many unwanted data; in addition, the language diversity used in Moroccan society adds many difficulties to the processing of tweets' content generated by Moroccan users.

C. Scikit-learn and Naive Bayes algorithm

Scikit-learn [18] is an open source library for machine learning that is simple and efficient for data mining and data analysis for the Python programming language. It is Built on *NumPy, SciPy,* and *Matplotlib* [10]; it includes many algorithms for classification, regression and clustering algorithm, and more. Because it is a robust library, we choose to Implement naive Bayes classifier in python with *scikit-learn*.

The Naive Bayes [19] is a supervised classification algorithm based on Bayes' Theorem with an assumption that the features of a class are unrelated, hence the word naive. The Naive Bayes classifier calculates the probabilities for every factor; then it selects the outcome with the highest probability.

Preprocessed tweets with NLP is given as input to train input set using Naïve Bayes classifier, then, trained model is applied to new collected tweets to generate either positive or negative sentiment.

The Bayes theorem is as follows:

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

Where:

- P(H): the probability of the hypothesis H being true. This is known as the prior probability.
- P(E): the probability of the evidence (regardless of the hypothesis).
- P(E|H) is the probability of the evidence given that hypothesis is true.
- P(H|E) is the probability of the hypothesis given that the evidence is there.

There are many applications of Naive Bayes Algorithms:

- Text classification/ Spam Filtering/ Sentiment Analysis
- Recommendation Systems.
- Real-time Prediction: Naive Bayes is a fast classifier, and it can be used for making predictions in real time
- Multi-class Prediction: more than two classes to be predicted.

D. PIG UDF

Apache Pig [8] is a popular data flow language; it is at the top of Hadoop and allows creating complex jobs to process large volumes of data quickly and efficiently. It will consume any data type: Structured, semi-structured or unstructured. Pig provides the standard data operations (filters, joins, ordering).

Pig provides a high-level language known as Pig Latin for programmers who are not so good at Java. It is a SQL-like language, which allows developers to perform MapReduce tasks efficiently and to develop their functions for processing data.

A Pig UDF [5] (User Defined Functions) is a function that is accessible to Pig but written in a language that is not PigLatin like Python, Jython or other programming languages; it is a function with a decorator that specifies the output schema.

We use Pig UDF to execute NLP program, written with Python language in a distributed manner using MapReduce. In consequence, the preprocessing became very fast and spread over the stored tweets.

E. Topic Modeling Using Latent Dirichlet Allocation(LDA)

Topic modeling allows us to organize, understand and summarize large collections of textual information. It helps to discover hidden topical patterns that are present in the collection; annotate documents according to these topics; and use these annotations to organize, search and summarize texts.

Topic models are unsupervised machine learning algorithms, which allow discovering hidden thematic structure in a collection of documents. These algorithms help us to develop new ways of text exploration. Many techniques are used to obtain topic models, but the most used one is Latent Dirichlet Allocation (LDA) [17].

LDA algorithm works as a statistical machine learning and text data mining; it allows discovering the different topics in a collection of documents. It consists of a Bayesian inference model that calculates the probability distribution over topics in each document, where each topic is characterized by a probabilistic distribution based on a set of words.

The LDA algorithm is used in our system, to discover the topics of classified tweets (positive and negative). For this reason, we implement a free python library for LDA called "Gensim" [20].

F. Interactive maps using Folium

Folium [11] is a powerful Python library that allows visualizing geospatial data onto interactive maps; it provides the facilities to transform coordinates to different map projections. The visualization happens "inline" or within the Python environment, using *IPython Notebook* and the results are interactive which makes this library very useful for dashboard building.

The Plotting of classified tweets in Moroccan map is necessary to discover the general mood in Moroccan regions as well as the dominant topics by using LDA.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. ARCHITECTURE OF THE SYSTEM

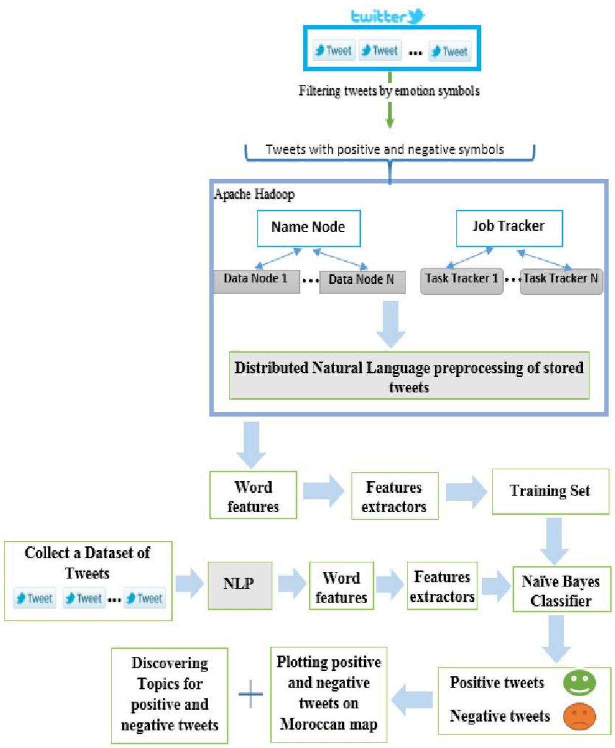


FIGURE 4. ARCHITECTURE OF THE SYSTEM

The first part of our system involves the extraction of data from the social network Twitter, which is an essential task in the data analysis process. All these tweets contain coordinates from different locations in Morocco and stored in the HDFS.

A. Data Description, Collection, and Storage

1) Data Description:

Twitter has unique conventions that make it distinct from other social network platforms; indeed tweets have many unique attributes, which differentiates Twitter from other platforms in the field of research. The First convention is the maximum length of a tweet, which is 140 characters, an average length of 14 words per tweet. The second convention is the availability of data using Twitter API; it is much easier to collect a large volume of tweets for training.

In Twitter messages, we can find acronyms, emoticons and other characters that express special meanings. The Emoticons can be represented using punctuation and letters or pictures; they express the user's mood. Emoticons can be categorized as:

• Happy emoticons (positive) :-) :) :D, etc

• Sad emoticons (Negative) :-(:(:-c, etc.

To use the method based on the emotion symbols we need to make an assumption. This assumption is that the emotion symbols in the tweet represent the overall sentiment contained in that tweet. This assumption is reasonable because in the majority of cases the emotion symbols will correctly represent the overall sentiment of that tweet as the maximum length of a tweet is 140 characters.

2) Data Collection:

Twitter Platform allows developers to collect data via Twitter API, but first, they need to create an account on *https://apps.twitter.com*. For each created account, Twitter provides four secret information: consumer key, consumer secret key, access token and access secret token, then we are authorized to access the database and retrieve tweets using the streaming API.

To get tweets that contain emoticons which are generated in Moroccan region, we filter tweets by location coordinates using the Streaming API and by a list of positive and negative emotion symbols. We get the geographical coordinates (latitude and longitude) of Morocco that we utilized in this filter, by using the specialized website in geolocation *http://boundingbox.klokantech.com*.

To handle the streaming of data from Twitter, we used Python library *Tweepy* [9] as shown in the script below that allows accessing to Twitter API.

```
import tweepy
import json
import hadoopy

consumer_key= XXX
consumer_secret= XXX
access_token= XXX
access_token_secret= XXX

hdfs_path_Pos = 'hdfs://master:54310/tweets_Positive/'
hdfs_path_Neg = 'hdfs://master:54310/tweets_Negative/'

POSITIVE = ["*O", "*-*", "*O*", "*o*", "* *",":P", ":D", ":d",
":p", ";P", ";D", ";d", ";p",":-)", ";-)", ":=)", ";=)", ":<)", ":>)",
";>)", ";=)", "=}", ":)", "(:;)","(;", ":}", "{:", ";}", "{;:]","[;",
":')", ";')", ":-3","{;", ":]",";-3", ":-x", ";-x", ":-X",";-X", ":-}", ";-
=}", ":-]",";-]", ":-.)","^_^", "^-^","😀",,"😁","😃", "😄","😅","
😆","😇","😈","😉","😊","😋","😌","😍","😎","😏","😗","😘
","😙","😚","😛","😜","😝"]

NEGATIVE = [":(", ";(", ":'(","=(", "={", "):", ");",")':", ")';", ")=",
"}=",";-{{", ";-{", ":-{{", ":-{",":-(", ";-(",":,)", ":'{","[:", ";]","😒
","😐","😑","😣","😞","😟","😠","😡","😥","😖","😢","
😦","😧","😨","😩","😪","😫","😭",,"😰","😱"]

class StdOutListener(tweepy.StreamListener):
  def on_data(self, data):

    decoded = json.loads(data)
    tweet_txt = decoded["text"]

    if decoded['place']['country_code'] == 'MA':
        …
```

```
        for tok in tweet_txt.split(" ")
            for emoji in POSITIVE:
                if emoji.decode('utf-8') in tok:
                hadoopy.put(localeFile ,hdfs_path_Pos)
                ...
stream.filter (locations=[-
17.2122302,21.3365321,0.9984289,36.0027875],async='true',enc
oding='utf8')
```

3) Data storage using HDFS

The storage of filtered tweets gathered from the Twitter API in HDFS is handled by using a Python wrapper for Hadoop called *Hadoopy* [6], which allows performing operations like reading and writing data from and to HDFS. We create Two folders in HDFS, one for the positive tweets (*'hdfs://master:54310/tweets_Positive/)'* and the other for the negative tweets (*'hdfs://master:54310/tweets_Negative/'*) as shown in the previous script.

B. Processing filtred tweets with NLP

A major issue which faces us when we are dealing with Twitter data is the informal style of the posts. Most tweets are written informally, contain many errors and abbreviations, and do not follow any grammatical rule. To minimize the effect of this informality on our classification, we will pre-process tweets, in order to clean them, before using them. We might find words misspelled, and therefore must be detected and corrected to evaluate sentiment more accurately.

Also, the linguistic diversity that characterizes the communication of Moroccan users on social network Twitter complicate the task of classification. To deal with this issue, we create a python file that contains a dictionary of words that we gathered manually, to transform words written in Moroccan dialect, or in a dialect of Berber Tamazight into Standard Arabic. These words could be written using the Arabic or French alphabet then we store it in each slave node of our cluster and imported inside the NLP script executed in these nodes. Below, a part of this file

```
#-*- coding: utf8 -*-
MoroccanDialects = [
("katbghi", u'تحب'),
("khas", u'يجب'),
("ban", u'يظهر'),
...
(u'إيييه', u'نعم'),
(u'عافاك' ,u'ارجوك'),
(u'خود' ,u'امسك'),
...
("zgizzi",u'يتجادل'),
("zigiz",u'يشغل'),
("werg ",u'يحلم ')]
```

Moroccan dialect written in French alphabet

Moroccan dialect written in Arabic alphabet

Moroccan Amazigh dialect "Tamazight"

The NLP step contains all the programs needed to preprocess the stored data, starting with parsing the tweets and extracting relevant information for our analysis, which are:

- *Text*: text of the tweet.
- *Lang*: language used by the user to write the tweet.
- *Coordinates:* location coordinates of a tweet.

The library used to preprocess tweets with NLP is the Natural language processing Toolkit (NLTK) [9], which is a set of open-source Python modules, allowing programs to work with the human language data. It involves capabilities for tokenizing, parsing, and identifying named entities as well as many more features; it also provides over 50 corpora and lexical resources such as WordNet and a set of text processing libraries.

We use the following steps for preprocessing the filtered tweets:

a) Delete unnecessary data: usernames, emails, hyperlinks, retweets, punctuation, possessives from a noun, duplicate characters, and special characters like smileys.

b) Shorten any elongated words (→ تكبير تكبيـــــــــــــــر)

c) Normalize whitespace (convert multiple consecutive whitespace characters into one whitespace character).

d) Convert hashtags into separate words, for example; the hashtag #SentimentAnalysis is converted into two words Sentiment and analysis.

e) Create a function to detect the language used to write the text of tweet (Standard Arab, French or English).

f) Create a function for automatic correction of spelling mistakes.

g) Create a list of contractions to normalize and expand words like (*What's=>What is)*

h) Delete the suffix of a word until we find the root. For example (*Stemming => stem)*

i) Remove tokens of part of speech that are not important to our analysis by using the Part-Of-Speech software of Stanford University. This software reads the text and assigns parts of speech (noun, verb, adjective) to each word.

j) Remove stopwords of standard Arabic *(أن, إن, يعد...)*, French (*alors, à, ainsi, ...*), and English *(about, above, almost, ...)*.

These steps are assembled in a python file called *NLTK_Tweet.py*. This file is executed in a distributed manner by an Apache Pig file called *Pig_Tweet.pig*. The file *NLTK_Tweet.py* needs to be registered in the script of the Pig file using *Streaming_python* as follows:

```
REGISTER 'hdfs://master:54310/apps/NLTK_Tweet.py' USING
streaming_python AS nltk_udfs;
```

The launch of our file *NLTK_tweet.py* is defined as follows:

```
data = LOAD '/tweets_Positive /*  using TextLoader() AS
(line:chararray);

Result =  FOREACH data GENERATE
nltk_udfs.NLTK_Function(line));
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C. Naïve Bayes Classifier

1) Data

Using Twitter API, we were able to collect experimentally a sample of 700 tweets (divided into positive and negative tweets) based on the emotion symbols and location filter, and 230 tweets as test set for accuracy evaluation of our classifier. All collected tweets are stored in a distributed manner using HDFS. The purpose of this paper, among others, is to be able to automatically classify a tweet as a positive or negative tweet. The classifier needs to be trained, that is why we use the stored tweets as training set after preprocessing step with NLP.

2) Implementation

For example, a fragment of the list of positive tweets looks like:

```
pos_tweets = [('I love this song, 'positive'),
              ('This picture is wonderful, 'positive'),
              ('I feel great this evening, 'positive'),
              ('This is my favorite food', 'positive')]
```

A fragment of the list of negative tweets looks like:

```
neg_tweets = [('I do not like this song, 'negative'),
              ('This picture is horrible', 'negative'),
              ('I feel sad this evening, 'negative'),
              (I hate this food, 'negative')]
```

We take these two lists and create a single list of tuples each containing two elements. The first element is an array containing the words and the second element is the type of sentiment. We ignore the words smaller than two characters, and we use lowercase for everything. The code is as follows:

```
tweets = []
for (words, sentiment) in pos_tweets + neg_tweets:
    words_filtered = [e.lower() for e in words.split() if len(e) >= 3]
    tweets.append((words_filtered, sentiment))
```

The tweets list now looks like this:

```
tweets = [
    (['love', 'this', 'song''], 'positive'),
    (['this', 'picture', 'wonderful'], 'positive'),
    (['feel', 'great', 'this', evening], 'positive'),
    (['this', 'favorite', 'food'], 'positive'),
    (['not', 'like', 'this', 'song'], 'negative'),
    (['this', picture, 'horrible'], 'negative'),
    (['feel', sad, 'this', evening], 'negative'),
   ([hate, 'this', 'food'], 'negative')])
]
```

3) Classifier

The list of word features needs to be extracted from the tweets. It is a list of every distinct word ordered by the frequency of occurrences. We use the following function and the two helper functions to get the list.

```
word_features = get_word_features(get_words_in_tweets(tweets))

def get_words_in_tweets(tweets):
    all_words = []
    for (words, sentiment) in tweets:
      all_words.extend(words)
    return all_words

def get_word_features(wordlist):
    wordlist = nltk.FreqDist(wordlist)
    word_features = wordlist.keys()
    return word_features
```

If we take a pick inside the function *get_word_features*, the variable 'wordlist' contains:

```
<FreqDist:
'this': 7,
'song': 2,
'feel': 2,
'evening': 2,
'picture': 2,
'wonderful': 1,
'favorite': 1,
'food':1
...
>
```

The list of word features is as follows:

```
word_features = [
'this',
'song',
'feel',
'evening',,
'picture',
'wonderful',
'favorite',
'food':1
    ...
]
```

The results show that 'this' is the most used word in our tweets, followed by 'song, then 'fell and so on …

We need to choose what features are pertinent to create our classifier. First, we need a feature extractor that returns a dictionary of words that are contained in the input passed. In our case, the input is the tweet. We use the word features list defined above along with the input to create the dictionary.

```
def extract_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains(%s)' % word] = (word in
document_words)
    return features
```

For example, let's call the feature extractor with the first positive tweet ['love', 'this', 'song']. We obtain the following dictionary which indicates that the document contains the words: 'love', 'this' and 'song'.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
{'contains(love)': True,
 'contains(evening)': False,
 'contains(this)': True,
 'contains(picture)': False,
 'contains(wonderful)': False,
 'contains(song)': True,
 'contains(favorite)': False,
 'contains(food)': False,
 'contains(horrible)': True,
 'contains(hate)': False,
 'contains(sad)': False,}
```

We use the method *apply_features* to apply the features to our classifier, and we pass the list of tweets along with the feature extractor defined above.

```
training_set = nltk.classify.apply_features(extract_features,
tweets)
```

The variable called 'training_set' contains the labeled feature sets, it is a list of tuples, where each tuple containing the feature dictionary and the sentiment category for each tweet.

```
[({'contains(love)': True,
   ...
   'contains(this)': True,
   ...
   'contains(song)': True,
   ...
   'contains(hate)': False,
   'contains(sad)': False},
  'positive'),
 ({'contains(love)': False,
   'contains(picture)': True,
   ...
   'contains(this)': True,
   ...
   'contains(wonderful)': True,
   ...
   'contains(hate)': False,
   'contains(sad)': False},
  'positive'),
 ...]
```

Now we can train our classifier using the training set.

```
classifier = nltk.NaiveBayesClassifier.train(training_set)
```

4) Testing the Classifier

To check the quality of our classifier by using the test set, we use an accuracy method in nltk that computes the accuracy rate of our model. Our approach reaches an accuracy of 69% which is considerate as a good value in our case.The simplest way to improving the accuracy of our classifier would be to increase the size of the training set.

```
import nltk.classify.util
print 'accuracy:', nltk.classify.util.accuracy(classifier, testTweets)
```

5) Classification of new collected tweets

Now that we have our classifier initialized and ready, we can try to classify collected and preprocessed tweets using NLTK and see what is the sentiment category output (positive or negative). Our classifier can detect that tweets have positive or negative sentiments. We evaluate our approach by streaming new collected tweets from Twitter API estimated at 300 tweets. A sample of collected tweets is as follows:

الوداد نسي ماتش الأهلي و بدأ يفكر في ماتش باتشوكا المكسيكي في كأس العالم و المصريين لازالو يحللون الهدف هل شرعي أو... https://t.co/PibiSBFoms from @monsef_filali at 11/07/2017 14:33

@lescitoyensorg Et ça continue ... from @cramounim at 11/07/2017 19:59

أداء مؤثر جدا و في قمة الخشوع لشيخي ماهر رعاه الله و حفظه مليكة @YouTube from @khadimarrahmane at 11/07/2017 15:39

Watching winner slowly realise that they're being kidnapped is the funniest thing ever #WinnerOverFlowers from @winneroediya at 11/07/2017 15:29

…

The below code is used to classify these new collected tweets using the classifier.

```
import nltk
from nltk.probability import FreqDist, ELEProbDist
from nltk.classify.util import apply_features,accuracy
...

print classifier.classify(extract_features(tweet.split()))
```

The output of the classification is the sentiment category of each tweet which is positive or negative. Our approach show good result despite the difficulties of multilingual tweets. some tweets are misclassified but we can override this issue by increasing the number of tweets in training set.

D. Topic Modeling with LDA

LDA is a probabilistic model used to determine the covered topics using the word frequency in the text. We use LDA in our approach for the classified tweets for each category(positive and negative). The LDA step will explain the reasons for the Moroccan user's mood. To generate the LDA model, we need to construct a document-term matrix with a package called "Gensim", which allows us to determine the number of occurrences of each word in each sentiment category. The LDA program used to discover topics is as follows:

```
from gensim import corpora, models
import hadoopy

fname_in = '/home /corpusTweetsSeniment.csv'
documentsPos = ""
documentsNeg = ""

with open(fname_in, 'rb') as fin:
    reader = csv.reader(fin)
    for row in reader:
        if row[3] == "positive":
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
        documentsPos = documentsPos + row[2] + ","
    elif row[3] == "negative":
        documentsNeg = documentsNeg + row[2] + ","

documentsPos =  documentsPos[:-1]
documentsNeg =  documentsNeg[:-1]

print ( ---- Topics  for positive tweets -----)
texts = [word.split() for word in documentsPos.split(",")]
dictionary = corpora.Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]
lda = models.ldamodel.LdaModel(corpus, id2word=dictionary,
num_topics=2 , passes=10)
lda.show_topics()

print ( ---- Topics  for negative tweets -----)
texts = [word.split() for word in documentsNeg.split(",")]
dictionary = corpora.Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]
lda = models.ldamodel.LdaModel(corpus, id2word=dictionary,
num_topics=2 , passes=10)
lda.show_topics())
```

For instance, the topics detected by our LDA model are:

```
---- Topics  for positive tweets ------
Topic #1: Maroc, football, equipe, russie, qualification
Topic #2: كأس، افريقيا، الوداد، جماهير

---- Topics  for negative tweets -----
Topic #1: تلوث، تهديد، سكان، بيئة، إشكالية
Topic #2: accident, circulation, mort, blesses, route
```

E. Plotting the classified tweets on map using Folium

During the streaming of filtered tweets from the Twitter API, we extract the coordinates (longitude and latitude) of each tweet. We then use these coordinates in Folium to show locations of tweets on our Moroccan map. The tweets that belong to positive mood are in green color and the negative mood are in red color. The developed program is as follows:

```
import folium
import csv

filename = '/home /corpusTweetsCoordinatesSeniment.csv'

map = folium.Map(location=[36.0027875,-17.2122302],
zoom_start=6)

  with open(filename) as f:
      reader = csv.reader(f)
      for row in reader:
        if row[3] == "positive":
          folium.Marker(location=[row[1],row[0]],popup=row[
          2],icon = folium.Icon(color ='green')).add_to(map)

        elif row[3] == "negative":
          folium.Marker(location=[row[1],row[0]],popup=row[
          2],icon = folium.Icon(color ='red')).add_to(map)

map.save("/home/abdeljalil/map_tweets.html")
map
```

The Figure 4 below shows the result of plotting classified tweets on the Moroccan map:

Figure 5. Locations of classified tweets on Moroccan map

This representation gives an idea about the locations of the Moroccan positive and negative tweets. This map and the topics generated by LDA are a good and perfect combination to study the mood of the Moroccan users, and more specifically to answer the two questions : Why this mood (LDA) and Where (Map).

V. Conclusion and Future Work

Twitter nowadays became one of the major tools and new types of the communication. People directly share their opinions through Twitter to the public. One of the very common analyses, which can be performed on a large number of tweets, is sentiment analysis. In the proposed work, we have presented a method for an automatic collection of a corpus that can be used to train a multilingual sentiment classifier so that it will be able to classify tweets into positive and negative. This classification is based on Naive Bayes classifier. Then we use methods to get insight from the classified tweets as the hidden topics and the locations of positive and negative tweets, which can conduct to better understanding of the Moroccan mood about different subjet and events. As future work, we plan to increase the accuracy of our classifier by increasing the number of filtered tweets, and by improving the preprocessing with NLP.

VI. References

[1] arabsocialmediareport, "Twitter in Arab Region". [Online]. Available: http://www.arabsocialmediareport.com/Twitter/LineChart.aspx. [Accessed: 01- Sep- 2017].

[2] Mrudula Varade and Vimla Jethani, "Distributed Metadata Management Scheme in HDFS", International Journal of Scientific and Research Publications,Volume 3, Issue 5, May 2013.

[3] M. Ghazi and D. Gangodkar, "Hadoop, MapReduce and HDFS: A Developers Perspective", Procedia Computer Science, vol. 48, 2015.

[4] M. Nagao, "Natural Language Processing and Knowledge", 2005 International Conference on Natural Language Processing and Knowledge Engineering.

[5] Pig.apache.org, "User Defined Functions". [Online]. Available: https://pig.apache.org/docs/r0.9.1/udf.html. [Accessed: 06- Sep -2017].

[6] "hadoopy", hadoopy.readthedocs.org. [Online]. Available: https://hadoopy.readthedocs.org/en/latest/. [Accessed: 01- Sep- 2017].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] "Apache Hadoop", [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop.html. [Accessed: 01- Mar- 2017]

[8] pig.apache.org, "Welcome To Apache Pig". [Online]. Available: https://pig.apache.org/. [Accessed: 01- Sep- 2017].

[9] nltk.org/, "nltk". [Online]. Available: http://www.nltk.org/. [Accessed: 01- Sep- 2017].

[10] matplotlib.org/basemap/, " Welcome to the Matplotlib Basemap Toolkit documentation". [Online]. Available: http://matplotlib.org/basemap/. [Accessed: 01- Sep- 2017].

[11] "Folium", [Online]. Available: https://github.com/python-visualization/folium. [Accessed: 01- Mar- 2017].

[12] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proc. of LREC, 2010.

[13] D. Davidov, O. Tsur and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys", Proceedings of Coling, 2010

[14] Jonathon Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", In ACL. The Association for Computer Linguistics, 2005.

[15] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval, 2008.

[16] Alec Go, Lei Huang, and Richa Bhayani, "Twitter sentiment analysis", Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group, 2009

[17] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research, 2003.

[18] scikit-learn.org/stable/, "scikit-learn". [Online]. Available: http://scikit-learn.org/stable/. [Accessed: 01- Sep- 2017].

[19] Z. Zhang, "Naïve Bayes classification in R", Annals of Translational Medicine, vol. 4, no. 12, pp. 241-241, 2016.

[20] radimrehurek.com/gensim/, "Gensim topic modeling for humans". [Online]. Available: https://radimrehurek.com/gensim/. [Accessed: 01- Sep- 2017].

[21] flume.apache.org/, "Apache Flume". [Online]. Available: https://flume.apache.org/. [Accessed: 01- Sep- 2017].

[22] oozie.apache.org/, " Apache Oozie Workflow Scheduler for Hadoop". [Online]. Available: http://oozie.apache.org/. [Accessed: 01- Sep- 2017].

[23] mahout.apache.org/, "Apache Mahout: Scalable machine learning and data mining". [Online]. Available: http://mahout.apache.org/. [Accessed: 01- Sep- 2017].

[24] avro.apache.org/, "Welcome to Apache Avro!". [Online]. Available: https://avro.apache.org/. [Accessed: 01- Sep- 2017].

[25] hbase.apache.org/, "Apache HBase – Apache HBase™ Home". [Online]. Available: https://hbase.apache.org/. [Accessed: 01- Sep- 2017].

[26] hive.apache.org/, "APACHE HIVE TM". [Online]. Available: https://hive.apache.org/. [Accessed: 01- Sep- 2017].

[27] spark.apache.org/, " Apache Spark™ - Lightning-Fast Cluster Computing". [Online]. Available: https://spark.apache.org/. [Accessed: 01- Sep- 2017].

Using biometric watermarking for video file protection based on chaotic principle

Saja J. Mohammed
Computer Science Department
College of Computer Science & Mathematics /Mosul University
Mosul/ IRAQ
Sj_alkado@yahoo.com

Abstract- **In the present time, security in the content of multimedia became one of significant science types. Watermarking is one type of multimedia protection, it is idea of protect digital components. Watermarking has extended and applied for many requirements, like fingerprinting, copyright protection, content indexing and many others watermarking application.**

The suggested algorithm is to hide a bio-watermarking encrypted data using video file as a cover in order to achieve video file protection. The recipient will need only to follow the required steps to retrieve the data of watermark. The idea of proposed method is based on hiding the watermark in audio partition of video file instead of video's image. Also use multiple frequency domains to hide the biometric watermark data using chaotic stream as key for encrypting the watermark and choose location for hiding. Subjective and objective tests (SNR, PSNR and MSE) are used to estimate the performance of the suggested method with applying simple attack that may attack the cover file.

Experimental result of the algorithm shows good recovering of watermark code which is virtually undetectable within the video file.

Keywords: video watermarking, DCT, DWT, Biometric system, chaotic.

I. INTRODUCTION

Nowadays, the digital media and the Internet have become so popular. That led to rise the requirements of secure data transmission. A number of useful techniques are proposed and already in use [1]. Watermark is one of these techniques which is a digital code embedded into the content of digital cover i.e. text, image , audio or video sequence [2].

Watermarking method is describe in the process as follows: Firstly, the abstraction of copyright data in the form of watermarks and imbedded in multimedia carriers using one of many embedding algorithms. After that, these carriers are distributed by the network or any digital storage. When necessary , the carriers are processed to detect the watermark existence . It is also possible to extract watermark for many various purposes[3].

In general, watermarking process is to embed some copyright data into the host data as an evidence ownership right. It must meet requirements which is: Security Obviously, Robustness, Imperceptibility and Capacity [4].

Various algorithms of digital video watermarking have been suggested. These techniques are categorized according to the domain which they working with. Some of these techniques embedded the watermark using the spatial domain by modification of the pixel values in each extracted video frame. These methods are entrusted to attacks and signal distortions. However, other techniques using the frequency domain to embed their watermark, this is the better robust to distortions[2].

Digital video is a sequence of still images merging with audio. The watermark will carry all types of information however the quantity of watermark data is limited. The vulnerability of the data is direct concerning of the amount of the information that carried by the watermark. The amount is absolutely limited by the size of particular video sequence[2].

II. WHAT IS BIOMETRICS?

Biometrics, is the process of authentication which depend on the physiological or behavioral properties and its ability to identify whether the person is authorized or not. Biometric properties distinctive as they can not be lost or forgotten, the presentation of identifying person will be done physically [5][6].

There are many of biometrics like fingerprint, face, hand thermogram, , signature, retina, iris, hand geometry, voice and so… .The most proven method is Iris -based identification . Iris can be defined as the colored part of eye, Fig. 1 shows the iris contents .The two eyes iris of any person have various iris pattern. Because the iris has a lot of characteristic which help to distinguish one iris from another, two conformable twins also have various iris patterns. Iris stills in a stable pattern not depended to the age affection that mean it stay in stability from the birth to the death. Also, the system of iris recognition can be un-invasive to their user[5][7].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

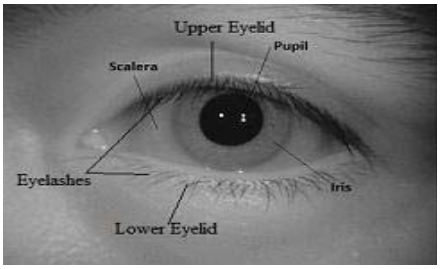


Figure 1. Structure of Iris

III. CHAOTIC SIGNAL

The chaotic signal is similar to noise signal, but it is certain in complete, that means if anyone has the initial values and the used function, that will be reproduce the same amount exactly. The profit of chaotic signal are[8]:

I. The initial conditions sensitivity

A minor variation in initial amount will cause important distinction in subsequent measures. The final signal will be differ completely if there is a small modification in the signal amount.

II. The accidental feature apparently

To compare with productive casual natural number in which the numbers scope cannot be generated again, the technique used for generating the same casual number in methods based on the chaotic function will create the ground that if the initial values and the used function are the same, the same number generated again.

III. The work deterministic

However, the chaotic functions were the casual manifest, they are wholly similar. That is if the initial values and the used function are fixed, the amounts of numbers will generate and re generate which seemingly have not any order and system. Logistic Map signal is one of the farthest known chaotic signals, this signal is presented by equation shown in (1):

$$X_{n+1} = rX_n (B\text{-}x_n) \qquad (1)$$

Where Xn gets the numbers in range [0,1]. The signal explain three various chaotic characteristics in three various ranges on the division of r parameter, the signal characteristics will be the best by assuming X0 =0.3.

- in $r \in [0,3]$, the signal characteristics in the first 10 iteration show some chaos and they were fixed after that , Fig. 2 (a)[9][10]
- in $r \in [3, 3.57]$, the signal characteristics in the first 20 iteration show some chaos , they were fixed after that, Fig. 2(b),
- in $r \in [3.57,4]$, the signal characteristics are chaotic in complete, Fig. 2(c)

Agreement with the above description and the requirements of the proposed algorithm to ensure complete chaotic characteristics for video watermarking, the logistic map chaotic signal with primary value X0=0.3 and $r \in [3.57,4]$ are used[9].

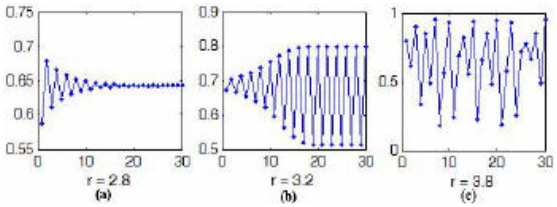


Figure. 2 The signal of logistic map chaotic with X0=0.3 and

$r \in [0,3]$, (b) $r \in [3, 3.57]$, (c) $r \in [3.57,4]$

IV. THE RELATED WORKS

There are many of watermarking methods based on video file as cover suggested in last period . One of these methods was proposed by Mobasseri (2000), who suggest a watermarking algorithm in compressed videos using spatial domain. Where Hong et al (2001) proposed an algorithm based on DWT they modify middle frequencies in the file. In other side Liu et al (2002) suggested a video watermarking algorithm using DWT to embed multi information bits. Chang & Tsai (2004) suggested a watermarking algorithm for a compressed video by VLC decoding and VLC code substitution. Zhong & Huang (2006) suggested video watermarking schema using spread-spectrum method for watermarking robustness improvement. Mirza et al (2007) suggest a video watermarking method using Principal Component Analysis [4].

V. THE PROPOSED METHOD

As we know video file format contain major two part of multimedia types: image and audio. It is generated by mixing the two kinds of multimedia types. The proposed method differs from the typical watermarking scheme. It is based on hiding watermark data in video's audio part instead of image one.

There are two categories of Digital watermarking technique: spatial domain watermarking technique and frequency domain watermarking techniques. The spatial domain methods hide the watermark using modifying some values of video file in directly way . The frequency domain technique will be embedding the watermark in best ways to ensure better determine of perception criterion and robust watermarking. Therefore the proposed algorithm used frequency domain to hide watermark data and in order to achieve more security multiple type of frequency domains with chaotic key are used.

In the proposed method, the watermark is based on biometrics (exactly on iris) to generate the watermarking code. The following sections discuss the proposed video Watermarking in details.

A) The proposed algorithm of embedding watermark code:

The proposed algorithm can be divided into two basic parts: generating the biometric watermark code and hiding it in video file data using chaotic key.

- *Generating the biometric watermarking code:*

To generate iris watermark data the iris (included in eye image) must be segmented .This will be made in the following steps : edge detection, circle detection and eyelid

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

detection. There are many technique for edge detection. This paper used canny edge detection and Hough transform to find iris and pupil boundaries. Iris image must be available in sender and receiver sides. For more security the watermark is encrypted using chaotic key.

The proposed algorithm of generating the bio-watermarking code is explained in the following steps:

Input: Iris image.

Output: Encrypted bio-watermarking code.

1) Begin
2) Choose iris image.
3) Apply iris segmentation.
4) Take iris data which is laying under pupil circle.
5) Apply edge detection using canny filter.
6) Generate chaotic key.
7) Encrypt iris data using the generated chaotic key.
8) End.

Fig. 3 shows the flowcharts of generating the bio-watermark code.

- *Embedding the watermark in video file using chaotic key:*

Input: Video file, Bio-watermark code.

Output: Watermarked video file.

1) Begin.
2) Choose video file to be cover file.
3) Split image and audio in it and consider audio part as a cover.

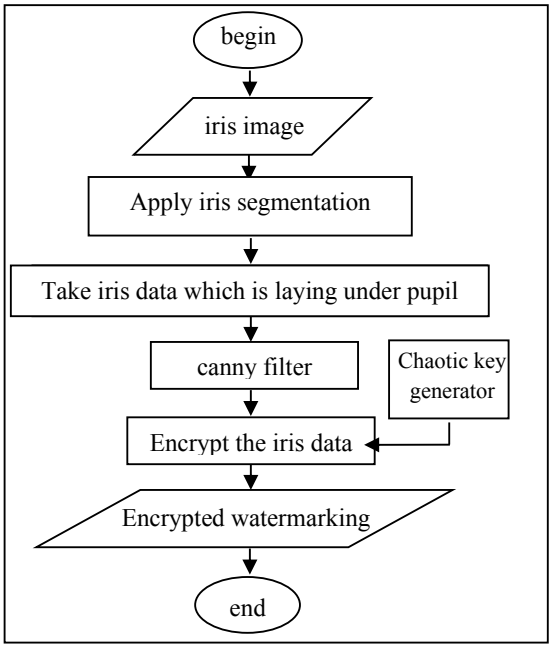


Figure 3. Generating the bio-watermarking code.

4) Apply DWT on audio part.
5) Apply DCT on resulted DWT coefficients.
6) Hide the length of watermark (Len) in first 4 bytes of cover data.
7) Generate chaotic key to be the index of chosen cover data .
8) Hide watermark code in cover by exchanging the fourth decimal number after comma in cover by another digit of watermark code.
9) Repeat this step until last digit in watermark code.
10) Apply DCT inverse, then DWT inverse.
11) Reformat the video cover.
12) End

Fig. 4 shows the proposed algorithm of hiding the biometric watermarking code in video file using chaotic key.

B) The proposed algorithm of extracting watermark code:

Input: The covered video file.

Output: Achieve video file protection or not.

1) Begin.
2) Input the covered video file.
3) Extract audio part from the covered video file.
4) Apply DWT on audio part.
5) Apply DCT on resulted DWT coefficients
6) Extract the length (Len) of watermark from first 4 byte in cover.
7) Generate chaotic key(for extracting and decryption operation).
8) Using the chaotic key to extract watermark code.
9) Repeat this step until reaching the length of watermark code.
10) Decrypt the extracted watermark using same chaotic key.
11) Independently… Generate the iris watermark code (origin one) by executing the steps of generating the biometric watermark (1 to 5).
12) Use the coparition between the onigin watermark with the extracted watermark data. If they are identical ,video file protection is achieved otherwise the file is not protected.
13) End

Fig.5 shows the proposed algorithm of extracting watermark code.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

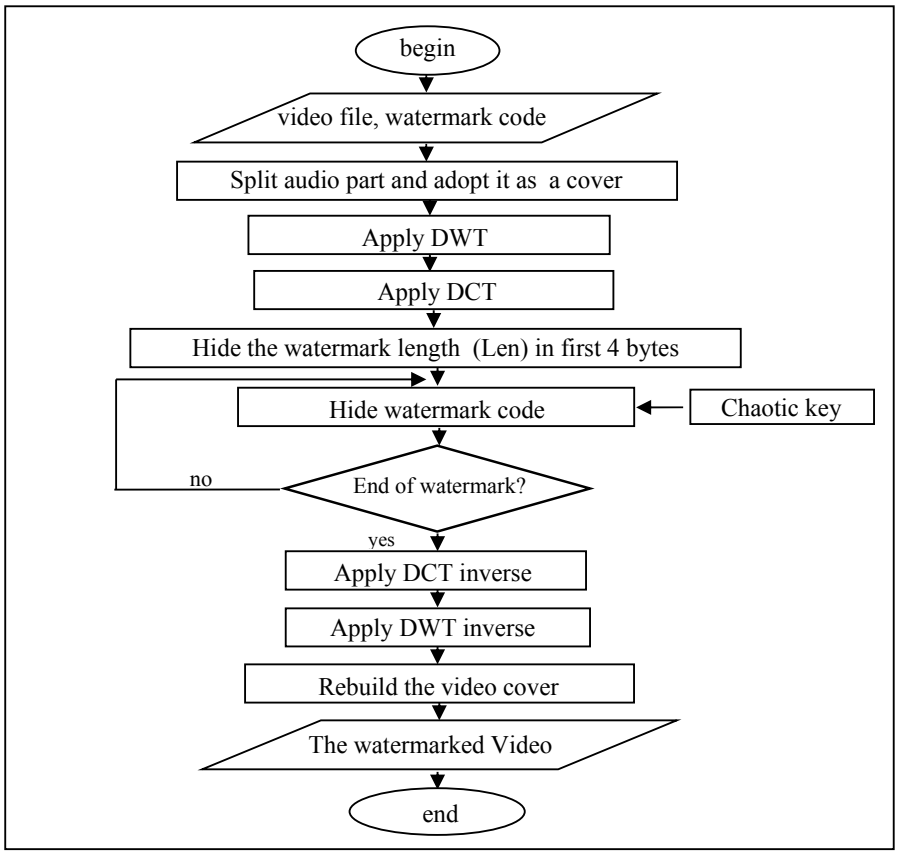


Figure. 4 The proposed algorithm of Hide the watermark in video file using chaotic key

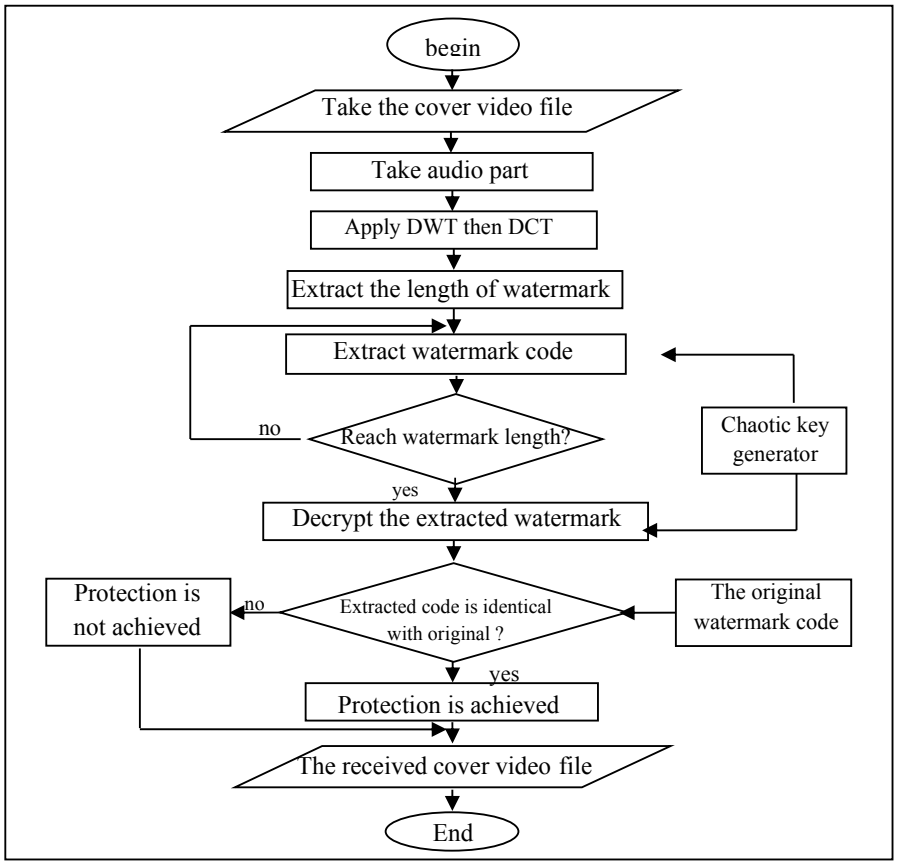


Figure 5. The proposed algorithm of extracting watermark code

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. EXPERIMENTAL APPLICATION AND RESULTS

A number of video sequences have been tested using the proposed method. The bio-watermark is extracted from the watermarked video and its robustness is checked by calculating some famous measures.

Moreover, the proposed method is applied on many iris images obtained from CASIA database. At last the iris code is obtained and hidden in video file. Figs 6,7,8 show the experimental steps that are done on iris image to get bio-watermark code.

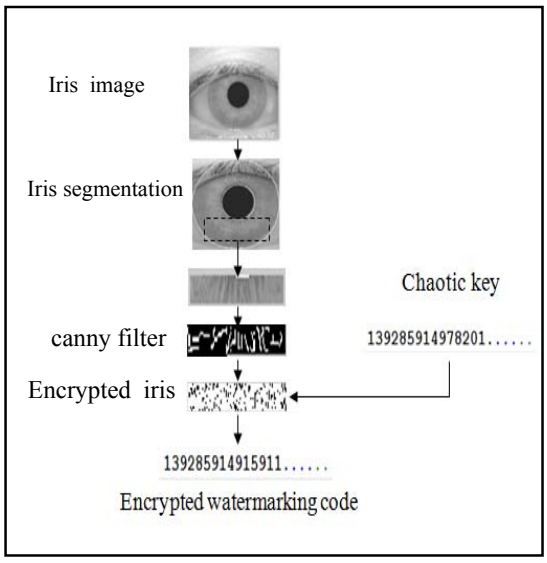


Figure 6. The proposed process for getting watermarking code

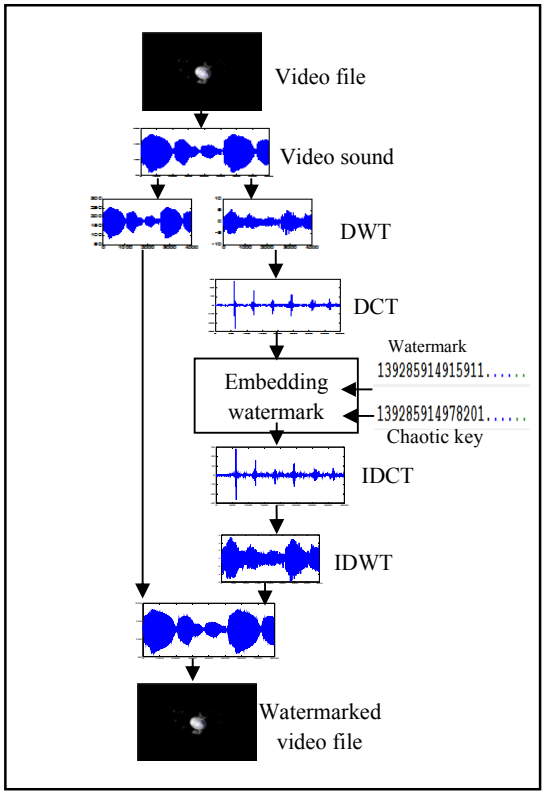


Figure 7. The proposed embedding process

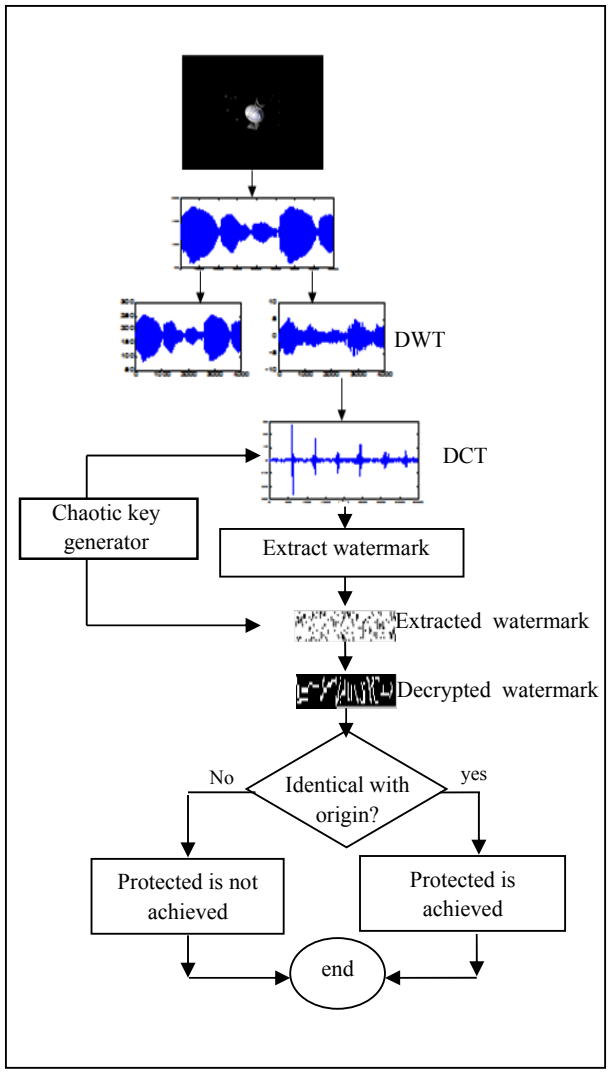


Figure 8. the proposed extracted process

A number of measures are applied on it to make sure that the proposed algorithm is strong enough to carry the watermark safely. Table I. explain the results of applying standard measures (Correlation, SNR,PSNR and MSE) to the proposed algorithm.

TABLE I. THE RESULTS OF APPLYING STANDARD MEASURES TO PROPOSED ALGORITHM

File name	Correlation	SNR	PSNR	MSE
Radar	1	219.3514	75.586	2.7631e-08
Morale	1	205.74	75.504	2.8152e-08
Test	1	212.03	75.826	2.6145e-08

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The watermarked video was attacked by simple types of watermarking attacks. This types of attacks are try to annoy the watermark by modify the whole cover without any attempt of identifying and separating the watermark [11][12]. Adding white noise (Gaussian noise) is applied to the video cover resulting from the proposed algorithm. Fig. 9 shows the effect of adding Gaussian noise to the video cover file with different signal to noise ratio values. While Table II. explains the output results of adding Gaussian noise to the video cover .

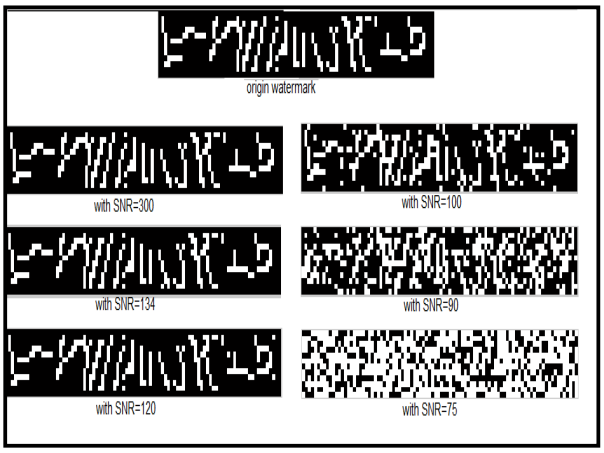


Figure 9. The effects of adding gaussian noise with variety values of signal to noise ratio

TABLE II. THE OUTPUT RESULT OF ADDING GAUSSIAN NOISE TO THE EMBEDDED WATERMARK

SNR	Correlation	MSE
200	1	0
150	1	0
134	0.8720	0.0743
120	0.7956	0.4149
100	0.1926	3.7147
90	0.0626	9.2799
75	0.0537	30.0978

VII. CONCLUSION

The paper propose an efficient method to embed a biometric watermarking in video file. It make use of two powerful mathematical transforms: DWT and DCT and applied them on the audio part of video file format instead of video's images. The proposed method use the chaotic sequence in order to find a video file locations in order to hide bio-watermark on the one hand and the sequence is used to encrypt and decrypt the bio-watermark data on the other.

After applying the proposed algorithm, some standard measures between the two watermarks(original and extracted one) are applied using correlation, SNR, PSNR and MSE. Also measures are applied on the attacked video file using correlation and MSE. The experimental results show their robustness against noise adding; very low noise watermark with expectable SNR values. The obtained results give to the proposed algorithm high performance with robustness in watermarking application in order to achieve protection to any video file.

REFERENCE

[1]. Bhaumik Arup , Choi Minkyu ,Robles Rosslin J. and Balitanes Maricel O. ," Data Hiding in Video", International Journal of Database Theory and Application, Vol. 2, No. 2, June 2009, p 9-16 .

[2]. Hood Ankita A. and Janwe N. J. ,"Robust Video Watermarking Techniques and Attacks on Watermark – A Review", International Journal of Computer Trends and Technology- volume4 Issue1 ,2013, p30-34.

[3]. Faragallah Osama S., "Efficient video watermarking based on singular value decomposition in the discrete wavelet transform domain", International Journal of Electronics and Communications (AE-) , Int. J. Electron. Commun. (AE-) 67 , 2013 , p189– 196 .

[4]. Bhatnagar Gaurav and Raman Balasubrmanian, "Wavelet packet transform-based robust video watermarking technique", Indian Academy of Sciences , Sadhana Vol. 37, Part 3, 2012, p 371–388.

[5]. Al-Gurairi Maha Abdul-Rhman Hasso," Biometric Identification Based on Improved Iris Recognition Techniques", A Ph. D. Thesis Submitted to The Council of the College of Computer and Mathematical Sciences, University of Mosul ,2006.

[6]. Waghmare L.M. and Roselin Vanaja, " Iris Texture Analysis for Security Systems" , International Journal of Computer Applications (0975 – 8887) Volume 64– No.22, 2013, p37-44.

[7]. Dhavale Sunita V. ," DWT and DCT based Robust Iris Feature Extraction and Recognition Algorithm for Biometric Personal Identification ", International Journal of Computer Applications (0975–8887) Volume 40– No.7, 2012, p 33-37.

[8] Enayatifar R. , Mahmoudi F. and Mirzaei K.," Using the chaotic map in image steganography. International Conference on Information Management and Engineering, 2009 ,p 491-495.

[9]. Saeed Melad J., " A New technique based on chaotic steganography and encryption text in DCT domain for color image", Journal of Engineering Science and Technology Vol. 8, No. 5 , 2013, p508 – 520 .

[10].Ahmed H.E., Kalash, H.M. and Farag Allah, O.S., "An efficient chaos-based feedback stream cipher (ECBFSC) for image encryption and decryption", Informatica, 31(1), 2007 ,p 121-129.

[11]. Ali Dujan Basheer Taha , "Digital Image Watermarking Techniques For Copyright Protection", A Ph. D. Thesis Submitted to The Council of the College of Computer Sciences & Mathematics , University of Mosul. , 2004.

[12].Zlomek Martin, " Video Watermarking", master thesis submitted to Department of Software and Computer Science Education, Charles University in Prague , Faculty of Mathematics and Physics, 2007.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Dynamic Cluster Formation Method To Increase The Lifetime Of A Wireless Sensor Network

K.Johny Elma
Assistant Professor, Department of IT
Jeppiaar SRR Engineering College
Chennai, India
kj_elma@rediffmail.com

Dr.S.Meenakshi
Professor and Head, Department of IT
Jeppiaar SRR Engineering College
Chennai, India
meenakshimagesh72@gmail.com

***Abstract* - A wireless sensor network consists of multiple detection stations called sensor nodes which has specialized transducers with a communication infrastructure for monitoring and recording physical and environmental conditions at diverse locations. Energy consumption of the network is crucial due to idle listening and overhearing. The sensor node's lifetime is the most critical parameter. The lifespan of a wireless sensor network is the total amount of time before the first sensor node runs out of power. An ideal cluster head is the one which has the highest residual energy. In the existing system, the cluster head loses its energy during data transmission and eventually becomes a dead node. Another node from the network is made as the cluster head. In the proposed system, we use Dynamic Cluster Formation Method to increase the lifetime of the network. In the proposed method, the clusters are formed dynamically based on its residual energy and the delay time. When the cluster head's energy drains to its threshold value, the cluster is again formed dynamically. Thus, the energy consumption is balanced by which the network lifetime is maximized.**

Keywords: Wireless Sensor Network; Sensor Node; Cluster; Residual Energy; Dead Node; Energy Consumption; Network Lifetime.

I. INTRODUCTION

A wireless sensor network (WSN) consists of sensor nodes capable of collecting information from the environment and communicating with each other via wireless transceivers. The sensor node is an autonomous small device that consists of mainly four units that are sensing, processing, communication and power supply. Sensor nodes have limited resources and it is difficult to deploy. Recharging the cluster nodes are even more difficult. Hence it is wise to use the available sensor nodes efficiently. These sensor nodes are deployed where human intervention is difficult. Hence collection of information is dependent on the sensor nodes. These sensors are used to collect the information from the environment and pass it on to base station. A base station provides a connection to the wired world where the collected data is processed, analyzed and presented to useful applications. Thus, by embedding processing and communication within the physical world, Wireless Sensor Network (WSN) can be used as a tool to bridge real and virtual environment. The collected data will be delivered to one or more sinks, generally via multi-hop communication. The sensor nodes are typically expected to operate with batteries and are often deployed to not-easily-accessible or hostile environment, sometimes in large quantities. It can be difficult or impossible to replace the batteries of the sensor nodes. Since multi-hop routing is generally needed, the nodes near a sink can be burdened with relaying a large amount of traffic from other nodes. A sensor node is a tiny device that includes four basic components. A sensing or actuating unit, a processing unit, transceiver unit and power supply unit [1, 2]. In addition to this, the sensor node may also be equipped with location detection unit such as a Global Positioning System (GPS), a mobilizer etc. Each sensing unit is responsible for gathering information from the environment as an input like temperature, pressure, light etc. and produces a related output in a form of electrical or optical signal. The analog signals produced by the sensor are converted to digital signals by the analog to digital communication (ADC) and fed into the processing unit. The transmitter and receiver are combined in to a single device called transceiver. Sensor nodes often use ISM (Industrial, Scientific and Medical) band. One of the most important components of a wireless sensor node is the power supply. The battery forms the heart of the sensor system as it decides the lifespan of the system. The battery lifespan needs to be prolonged to maximize the network lifespan. Small size of a sensor node results in corresponding constraints on memory also. Sensor nodes have very simple memory architecture.

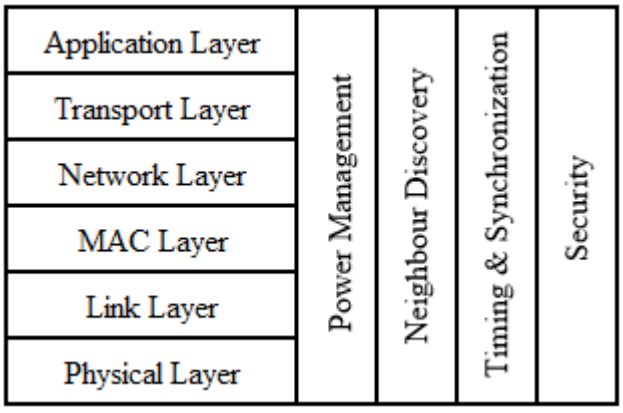


Figure 1. Protocol Stack

Sensor nodes use flash memories due to their cost and storage capacity. The mostly used operating system in sensor node are tiny OS (Operating System) Sensor nodes are resource

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

constrained in terms of energy, processor, memory, low range communication and bandwidth. Prolonging network lifetime is a critical issue. Thus, a good WSN design needs to be energy efficient. Energy consumption of one sensor node is influenced by the structure of protocol layers and the way each layer manages the sensing data.

II. RELATED WORKS

S. D. Muruganathan *et.al*, "A centralized energy-efficient routing protocol for wireless sensor network" had described the wireless sensor network and design issues [20]. They have proposed centralized routing protocol called base-station controlled dynamic clustering protocol (BCDCP), which distributes the energy dissipation evenly among all sensor nodes to improve network lifetime and average energy savings. The performance of BCDCP is then compared to clustering-based schemes such as low-energy adaptive clustering hierarchy (LEACH), LEACH-centralized (LEACH-C), and power-efficient gathering in sensor information systems (PEGASIS). Simulation results show that BCDCP reduces overall energy consumption and improves network lifetime over its comparatives.

O.B. Akan *et.al*, "Event-to-sink reliable transport in wireless sensor networks" had proposed a new reliable transport scheme for WSN, the event-to-sink reliable transport (ESRT) protocol, is presented in this paper. ESRT is a novel transport solution developed to achieve reliable event detection in WSN with minimum energy expenditure. It includes a congestion control component that serves the dual purpose of achieving reliability and conserving energy [18]. Importantly, the algorithms of ESRT mainly run on the sink, with minimal functionality required at resource constrained sensor nodes. ESRT protocol operation is determined by the current network state based on the reliability achieved and congestion condition in the network. This self-configuring nature of ESRT makes it robust to random, dynamic topology in WSN. Furthermore, ESRT can also accommodate multiple concurrent event occurrences in a wireless sensor field. Analytical performance evaluation and simulation results show that ESRT converges to the desired reliability with minimum energy expenditure, starting from any initial network state.

Wei-Peng Chen *et.al*, "Dynamic clustering for acoustic target tracking in wireless sensor networks" had devised and evaluated a fully decentralized, light-weight, dynamic clustering algorithm for target tracking. Instead of assuming the same role for all the sensors, we envision a hierarchical sensor network that is composed of 1) a static backbone of sparsely placed high-capability sensors which assume the role of a cluster head (CH) upon triggered by certain signal events and 2) moderately to densely populated low-end sensors whose function is to provide sensor information to CHs upon request. A cluster is formed and a CH becomes active, when the acoustic signal strength detected by the CH exceeds a predetermined threshold. The active CH then broadcasts an information solicitation packet, asking sensors in its vicinity to join the cluster and provide their sensing information. Through both probabilistic analysis and ns-2 simulation, we use with the use of Voronoi diagram, the CH that is usually closes to the target is (implicitly) selected as the leader and that the proposed dynamic clustering algorithm effectively eliminates contention among sensors and renders more accurate estimates of target locations as a result of better quality data collected and less collision incurred [27].

Weifa Liang *et.al*, "Online data gathering for maximizing network lifetime in sensor networks" had considered an online data gathering problem in sensor networks, which is stated as follows: assume that there is a sequence of data gathering queries, which arrive one by one. To respond to each query as it arrives, the system builds a routing tree for it. Within the tree, the volume of the data transmitted by each internal node depends on not only the volume of sensed data by the node itself, but also the volume of data received from its children. The objective is to maximize the network lifetime without any knowledge of future query arrivals and generation rates. In other words, the objective is to maximize the number of data gathering queries answered until the first node in the network fails. We then show the problem to be NP-complete and propose several heuristic algorithms for it. We finally conduct experiments by simulation to evaluate the performance of the proposed algorithms in terms of network lifetime delivered [26]. The experimental results show that, among the proposed algorithms, one algorithm that takes into account both the residual energy and the volume of data at each sensor node significantly outperforms the others.

Yong Yuan *et.al*, "Virtual mimo-based cross-layer design for wireless sensor networks" A novel multi-hop virtual multiple-input-multiple-output (MIMO) communication protocol is proposed by the cross-layer design to jointly improve the energy efficiency, reliability, and end-to-end (ETE) QoS provisioning in wireless sensor network (WSN)[28]. In the protocol, the traditional low-energy adaptive clustering hierarchy protocol is extended by incorporating the cooperative MIMO communication, multi-hop routing, and hop-by-hop recovery schemes. Based on the protocol, the overall energy consumption per packet transmission is modeled and the optimal set of transmission parameters is found. Then, the issues of ETE QoS provisioning of the protocol are considered. The ETE latency and throughput of the protocol are modeled in terms of the bit-error-rate (BER) performance of each link. Then, a nonlinear constrained programming model is developed to find the optimal BER performance of each link to meet the ETE QoS requirements with a minimum energy consumption. The particle swarm optimization (PSO) algorithm is employed to solve the problem. Simulation results show the effectiveness of the proposed protocol in energy saving and QoS provisioning.

III. SYSTEM ANALYSIS

A. Existing Model

Sensor nodes are resource constrained in term of energy, processor, memory, low range communication and bandwidth. Limited battery power is used to operate the sensor nodes and is very difficult to replace or recharge it, when the nodes die. This will affect the network performance. Energy conservation increases lifetime of the network. Wireless sensor networks

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

consist of battery-powered nodes that are endowed with a multitude of sensing modalities including multi-media (*e.g.*, video, audio) and scalar data (*e.g.*, temperature, pressure, light, magnetometer, infrared). Although there have been significant improvements in processor design and computing, advances in battery technology still lag, making energy resource the fundamental challenge in wireless sensor networks. Consequently, there have been active research efforts on performance limits of wireless sensor networks. Those operations for a sensor to consume energy are target detection, data transmission and reception, data processing, etc. Among others data transmission consumes most of the energy, and it heavily depends on the transmission distance and the amount of transmitted data.

When the data transmission occurs, the energy of the cluster head drains and eventually dies. The lifespan of a wireless sensor network is the total amount of time before the first sensor node runs out of power. LEACH is dependent on the probability model. Some cluster heads may be very close to each other. These disorganized cluster heads could minimize the energy efficiency. To overcome the defects of LEACH methodology, a cluster head selection method, High Energy First (HEF) algorithm has been introduced. This method proves that the network lifetime can be efficiently increased. For mission critical WSN applications, it is important to be aware of whether all sensors can meet their mandatory network lifetime requirements. The High Energy First (HEF) algorithm is proven to be an optimal cluster head selection algorithm that maximizes a hard N-of-N lifetime for HC-WSNs under the ICOH condition. But lifetime of the network is much lesser when compared with the proposed system.

B. *Proposed Model*

The wireless sensor network (WSN) is partitioned into several clusters based on the coverage and connectivity. First, the coverage range is checked by all the nodes in a network. This is done by broadcasting a message to all its neighbor nodes. The nodes in the sensing range send an update message to that particular node. The node which receives maximum number of messages as reply becomes a cluster head (CH). A cluster is formed based on the chosen cluster head (CH). Data transmission occurs via cluster head (CH) which means all the nodes in a cluster send their data first to the cluster head which is then passed on to the base station. From the base station, the data is being sent to the receiver. The proposed method for the project is Dynamic Cluster Formation Method (DCFM).

There are two important parameters involved in DCFM. They are the residual energy and the delay time. The node with the minimum delay time and maximum residual energy is made the cluster head. A threshold value for the energy is maintained. When the cluster head's energy drains to its threshold value, a new cluster head chosen based on its residual energy. Again, the nodes in the cluster broadcast a message to all its neighbor nodes. The nodes which are in the sensing range sends an update message to that particular node. This is done to use the energy of nodes efficiently. The cluster is again formed dynamically. Thus, the energy consumption is balanced by which the network lifetime is maximized.

IV. SYSTEM ARCHITECTURE AND PROTOCOL DESIGN

A. *System Architecture*

Dynamic Cluster Formation Method involves the formation of clusters in a network dynamically. This method is mainly used to increase the lifetime of a network which is less in the existing system. Initially all the nodes are deployed in a network. In order to divide it into clusters, the cluster head is selected.

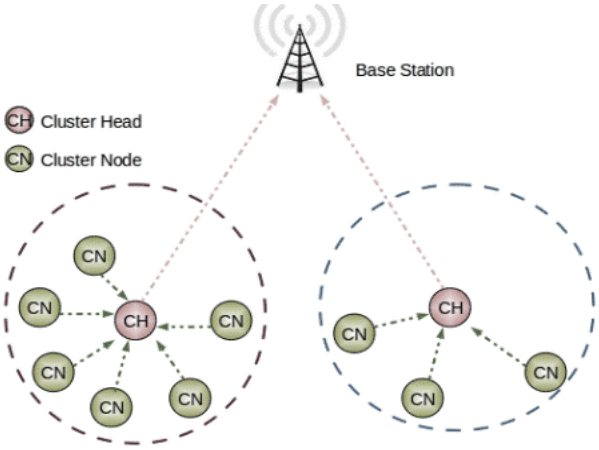


Figure 2. General Architecture of the Proposed System

The cluster head selection is dependent upon the residual energy and the delay time. The nodes which are in the sensing range of the cluster head groups to form a cluster. Initially a broadcast message is sent by all the nodes to all the other nodes. Thus, the number of broadcast messages a particular node receives is determined as node count. From the node count, the delay time is calculated. The node with minimum delay time and maximum residual energy is made the cluster head. From the cluster head the cluster is formed.

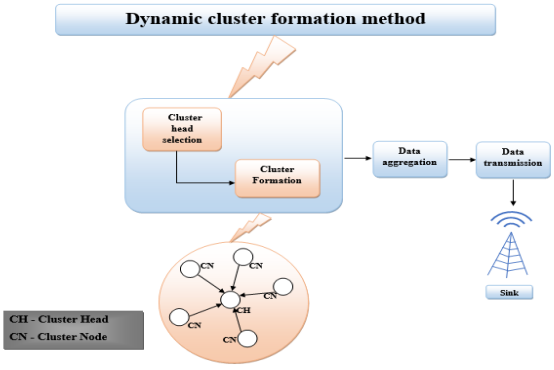


Figure 3. Proposed System Architecture

A threshold value is set for the energy of a node. During the data transmission, the energy of all the nodes drains. When

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the energy of the cluster head (CH) drains to the threshold value, another node is made the cluster head based on the residual energy and the delay time. Thus, the clusters are again formed dynamically by changing the cluster heads (CH). The cluster node (CN) senses the data and the sensed data is sent to the cluster head (CH). The cluster head (CH) receives all the data from all the nodes and does data aggregation. Once all the data are aggregated, the cluster head (CH) sends the data to the base station (BS). From the base station, the data is moved to the corresponding cluster's cluster head (CH). The cluster head (CH) transmits the data to the corresponding cluster node. Thus, the lifetime of the network is increased as there is a uniform consumption of energy in the network.

B. Protocol Design

The number of nodes is a network is denoted by the parameter 'n'. Each node has an initial energy $E_i(x)$, data transmission power $P_{tx}(x)$ and data reception power $P_{rx}(x)$. The cluster head CH is responsible for the transmission of the data from a particular cluster to other nodes. The selection of cluster head CH is dependent on the delay time $T_{delay}(x)$ and the residual energy $E_{res}(x)$. The node with the highest residual energy and lowest delay time $T_{delay\text{-}min}(x)$ becomes the cluster head CH.

Step 1: Deploy all the nodes in a network.

Step 2: For each node x, assign the initial energy $E_i(x)$, data transmission power $P_{tx}(x)$ data reception power $P_{rx}(x)$ and transmission range.

Step 3: A broadcast message is sent by the node x to all the other nodes which are in the sensing range of the node x. The message is represented as **(bcm, x).** The number of broadcast messages a particular node receives Nc(x) is determined. Here, Nc(x) is the node count of the node x.

Step 4: Calculate the delay time for the node x with node count Nc(x) as input. The delay time is given as,

$$\mathbf{T_{delay}(x) = C\ e^{1/Nc(x)}} \qquad \mathbf{(1)}$$

where C is a constant.

Step 5: The steps are repeated for all the nodes.

Step 6: The node with the lowest delay time $T_{delay\text{-}min}$ is determined from the delay time of all the nodes. An update message is sent by the node with lowest delay time to all nodes under its sensing range that it is the cluster head CH and forms a cluster. The message is represented as **(upm, x)**. If the delay time $T_{delay}(x)$ is same for more than one node, the node with the highest residual energy $E_{res}(x)$ is made the cluster head CH.

C. Energy Calculation

The energy of a node drains whenever there is a transmission or reception of data. When the energy of a cluster head CH drains to a threshold level (Thresh), another node is made the cluster head CH by following the above steps. The energy consumption of a node $E_{cmp}(x)$ is determined by the formula

$$\mathbf{Ecmp(x) = [Ptx(x) * N(tx)] + [Prx(x)*N(rx)]} \qquad \mathbf{(2)}$$

where $P_{tx}(x)$ is the data transmission power,
$P_{rx}(x)$ is the data reception power,
N(tx) is the number of transmissions,
N(rx) is the number of receptions.

The residual energy of a node $E_{res}(x)$ is determined by using the initial energy of the node $E_i(x)$ and the energy consumption of the node $E_{cmp}(x)$. It is given as,

$$\mathbf{Eres(x) = Ei(x) - Ecmp(x)} \qquad \mathbf{(3)}$$

Based upon this residual energy, the node with the maximum $E_{res}(x)$ is selected as the cluster head. The calculated residual energy is used in the selection of the cluster head. The node with the maximum residual energy and a minimum delay time is selected as the cluster head.

V. METHODLOGY

A. Dynamic Cluster Formation Technique

The sensor nodes are randomly distributed in a heterogeneous environment. The formation of cluster and energy efficient routing is done by the Dynamic Cluster Formation Method (DCFM).

- Cluster Formation: The sensor nodes are spatially distributed autonomous devices which are used for sensing, processing and communication purposes. All these nodes must be divided into clusters. Initially a network is divided into fields and the nodes are deployed in the network. The nodes which are used here are the sensor nodes which performs some processing, gathering information and communicating with each other. In order to communicate with each other, the nodes need to form a cluster. Thus, the formation of cluster is dependent upon the node's sensing range. The nodes which are in the sensing range of the cluster head becomes a member of the cluster. If a node is in the sensing range of one or more cluster, it becomes a member of the cluster which senses it first. Thus, all the nodes in a network will be a member of any one cluster of that network.
- Cluster Head Selection: Once all the nodes are deployed, it is necessary to form clusters. The formation of clusters helps in the communication of sensor nodes. The formation of cluster is basically dependent on the cluster head (CH). Hence the cluster head selection is the most important part. There are two parameters which are important in the cluster head selection. The delay time and the energy of the nodes decide which node must become a cluster head (CH). Initially, all the nodes in a network sends a broadcast message to all the other neighbor nodes. The number of message a particular node receives is determined as the node count. The delay time is calculated based upon its node count. Thus, the node with maximum energy and minimum delay time is made the cluster head. Cluster heads (CH)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

helps in the formation of clusters. Once the node with minimum delay and maximum energy is selected, the node sends an update message to all its neighbor nodes that it is the cluster head. Thus, all the nodes which receives the update message becomes a part of that particular cluster.

```
INPUT: Transmission_Power Ptx
       Reception_Power Rtx
       Initial_Energy Ei
BEGIN:
For_each_node(Current_node C(x))
{
    //Initialize Ei(x) = Ei;
    For_each_round(Current_trip C(r))
    {
        //Calculate Node_Count Nc(x);
        Calculate Transmission_Count N(tx);
        Calculate Reception_Count N(rx);
        Calculate Energy_Consumption_Of_Node Ecmp(x) with N(tx), N(rx), Ptx, Prx;
        Calculate Residual_Energy Eres(x) with Ei(x), Ecmp(x);
        //Calculate Delay_Time Tdelay(x);
    }
    Update Ei(x) with Eres(x);
}
```

The number of nodes is a network is denoted by the parameter 'n'. Each node has an initial energy Ei(x), data transmission power Ptx(x) and data reception power Prx(x). The cluster head CH is responsible for the transmission of the data from a particular cluster to other nodes. The selection of cluster head CH is dependent on the delay time Tdelay(x) and the residual energy Eres(x). The node with the highest residual energy and lowest delay time Tdelay-min(x) becomes the cluster head CH. Deploy all the nodes in a network initially, For each node x, assign the initial energy Ei(x), data transmission power Ptx(x), data reception power Prx(x) and transmission range. A broadcast message is sent by the node x to all the other nodes which are in the sensing range of the node x. The message is represented as (bcm, x). The number of broadcast messages a particular node receives Nc(x) is determined. Here, Nc(x) is the node count of the node x. Calculate the delay time for the node x with node count Nc(x) as input. The steps are repeated for all the nodes. The node with the lowest delay time Tdelay-min is determined from the delay time of all the nodes. An update message is sent by the node with lowest delay time to all nodes under its sensing range that it is the cluster head CH and forms a cluster. The message is represented as (upm, x). If the delay time Tdelay(x) is same for more than one node, the node with the highest residual energy Eres(x) is made the cluster head CH.The residual energy of a node Eres(x) is determined by using the initial energy of the node Ei(x) and the energy consumption of the node Ecmp(x).Based upon this residual energy, the node with the maximum Eres(x) is selected as the cluster head. Thus, the cluster head is dynamically selected using Dynamic Cluster Formation Method (DCFM).

- Dynamic Cluster Formation: Initially, the clusters are formed based upon the delay time and energy. Thus, the data transmission is involved where by which the energy of the nodes reduces. The energy is lost during the data transmission as well in the data reception. All the nodes in a cluster sends their data only via their cluster head. Thus, the cluster head loses more amount of energy. In the existing system, the energy drains completely and the node eventually dies. But this is the major disadvantage which is present in the existing system. The network lifetime is also very less. In order to overcome this disadvantage, we propose a method called Dynamic Cluster Formation Method (DCFM). By following this method, the network lifetime is increased. We use a threshold value for the energy of the node. This threshold value is used to balance the energy of the nodes in a network. The energy of the node decreases in different phases. They lose their energy during the data transmission and also in the data reception. Since cluster head is involved in the data aggregation, it receives the data from all the nodes. Thus, there would be a greater loss of energy in the cluster head. When the energy of the cluster head drains to the threshold value, a new node is made the cluster head based on DCFM. This is done by choosing a node which has maximum residual energy and minimum delay time. The same process is repeated on the loss of energy at the threshold level. For calculating the residual energy, the energy consumption is calculated. The energy consumption is calculated by taking into account the data transmission power Ptx(x), data reception power Prx(x), number of transmissions N(tx) and number of reception N(rx). Thus, by multiplying the number of transmissions and the transmission power along with the number of receptions and the reception power, the energy consumption is being determined. From the energy consumption, the residual energy is determined. Thus, the clusters are formed dynamically with the help of their residual energy.
- Energy Efficient Routing: In contrast to simply establishing correct and efficient routes between pair of nodes, one important goal of a routing protocol is to keep the network functioning as long as possible. The goal can be accomplished by minimizing cluster node's (CN) energy not only during active communication but also when they are inactive. Transmission power control and load distribution are two approaches to minimize the active communication energy, and sleep/power-down mode is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

used to minimize energy during inactivity. The parameters which involves in energy consumption include,

- Time to partition a network,
- Variance in node power levels,
- Cost/packet
- Maximum node cost.

The first metric is useful to provide the min-power path through which the overall energy consumption for delivering a packet is minimized. Here, each wireless link is annotated with the link cost in terms of transmission energy over the link and the min-power path is the one that minimizes the sum of the link costs along the path. However, a routing algorithm using this metric may result in unbalanced energy spending among mobile nodes. When some nodes are unfairly burdened to support many packet-relaying functions, they consume more battery energy and stop running earlier than other nodes disrupting the overall functionality of the ad hoc network. Thus, maximizing the network lifetime (the second metric shown above) is a more fundamental goal of an energy efficient routing algorithm: Given alternative routing paths, select the one that will result in the longest network operation time. The routing protocol that is used here is Ad-hoc on-demand Distance Vector Routing (AODV).

- Adhoc On-Demand Distance Vector Routing: The reactive on demand routing protocols establish the route to a particular destination only if it is needed. Adhoc on-demand Distance Vector (AODV) is one of the commonly used reactive on demand routing protocols in mobile ad hoc network (MANET). AODV is a reactive enhancement of the DSDV protocol. The route discovery process involves ROUTE REQUEST (RREQ) and ROUTE REPLY (RREP) packets. The source node initiates the route requested through the route discovery process using RREQ packets. The generated route request is forwarded to the neighbors of the source node and this process is repeated till it reaches the destination. On receiving a RREQ packet, an intermediate node with route to destination, it generates a RREP containing the number of hops required to reach the destination. All intermediate nodes that participates in relaying this reply to the source node creates a forward route to destination. AODV minimizes the number of packets involved in route discovery by establishing routes on-demand. The sample15.tcl shows a node configuration for a wireless mobile node that runs AODV as its adhoc routing protocol. Prior to the establishment of communication between the source and receiver node, the routing protocol should be mentioned to find the route between them. Data Transmission is established between nodes using UDP agent and CBR traffic.

VI. RESULTS AND DISCUSSION

The main aim of the project is to improve the network lifetime. A cluster is collection of nodes and in this case, sensor nodes are grouped to form a cluster in a network. This is done by choosing the cluster head dynamically using the method DCFM. It is seen that the energy is uniformly utilized and the network lifetime is increased when compared with that of High Energy First (HEF) algorithm. It is graphically represented using XGraph. To analyze a particular behavior of the network, users can extract a relevant subset of text-based data and transform it to a more conceivable presentation. Thus it is proven that the network lifetime is increased.

- Deployment of Nodes: Wireless Sensor Network consists of multiple sensor nodes. All the nodes are deployed in the network in such a way that the nodes can communicate with each other. The nodes senses the information and sends it to the base station. The nodes communicate with only their neighboring nodes.

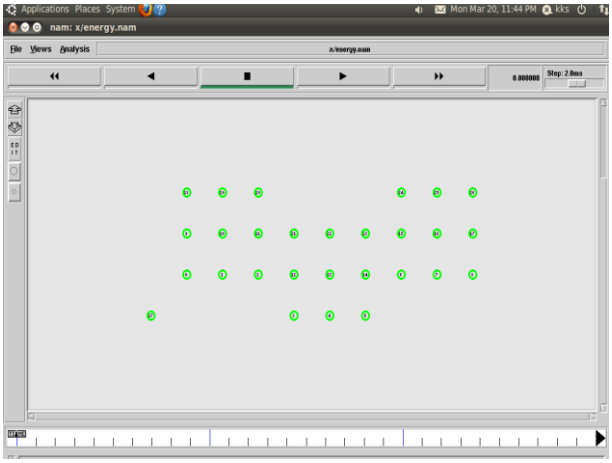
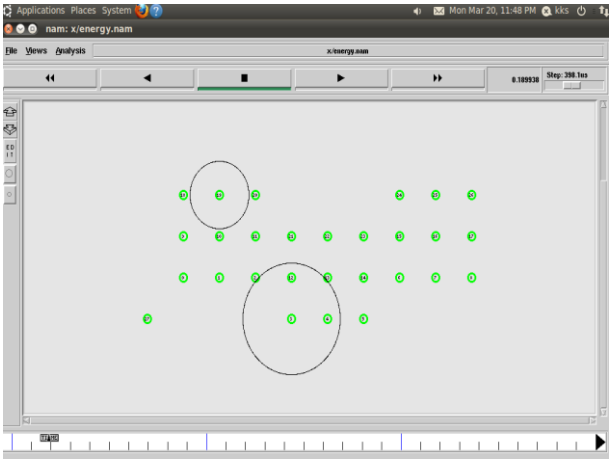


Figure 4. Deployment of Nodes

- Coverage Sensing: Identification of the nodes which are surrounding a particular node is done by sensing. A node does coverage sensing to find out its neighbor nodes for communicating. The nodes senses and sends it to the neighboring nodes and then finally it reaches the base station.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 5. Coverage Sensing

- Cluster Formation: After the coverage is sensed, all the nodes in a network will be a member of a cluster. Thus various clusters are being formed according to their sensing range. Now the transmission of data is done through the cluster head. A node is made the cluster head based upon the energy and delay time. All the nodes in a cluster transmits the data via the corresponding cluster head. The cluster head aggregates the data and passes it to the base station.

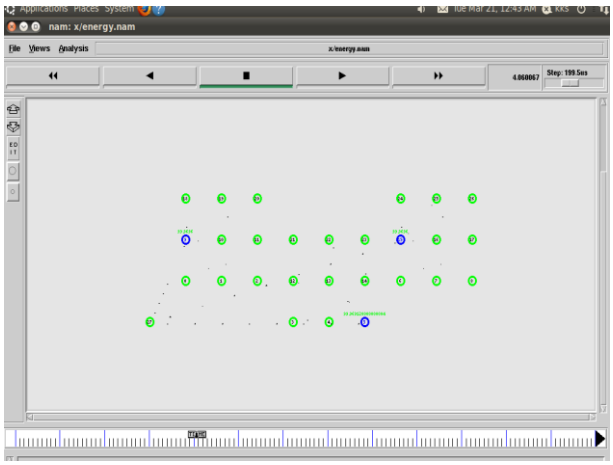


Figure 6. Formation of Clusters

- Cluster Head Selection: The nodes in a network sends a broadcast message to all the other neighbor nodes. The number of message a particular node receives is determined as the node count. The delay time is calculated based upon its node count. Thus, the node with maximum energy and minimum delay time is made the cluster head. Cluster heads (CH) helps in the formation of clusters. Once the node with minimum delay and maximum energy is selected, the node sends an update message to all its neighbor nodes that it is the cluster head. Thus, all the nodes which receives the update message becomes a part of that particular cluster.

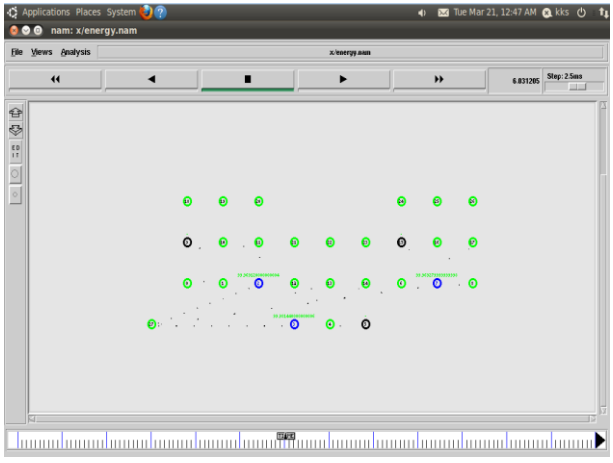


Figure 7. Cluster Head Selection

- Dynamic Clusteer Formation: All the nodes in a cluster sends their data only via their cluster head. Thus, the cluster head loses more amount of energy. We use a threshold value for the energy of the node. This threshold value is used to balance the energy of the nodes in a network. . When the energy of the cluster head drains to the threshold value, a new node is made the cluster head. Thus the node with maximum residual energy is made the cluster head.

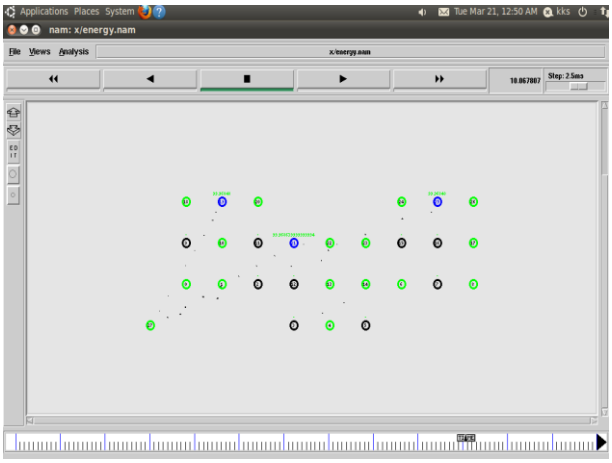


Figure 8. Dynamic Cluster Formation

Thus the energy is efficiently utilized and the lifetime of the network increases efficiently. Our experiment results show that the Dynamic Cluster Formation Method (DCFM) achieves significant performance improvement over High Energy First (HEF) algorithm, and DCFM's lifetime can be bounded.

VII. PERFORMANCE EVALUATION

The lifetime of a network is generally defined as the duration from the start to when the percentage of dead nodes comes to a threshold. It is seen that the lifetime of the network is less in High Energy First (HEF) algorithm. When compared with DCFM, DCFM provides a better network lifetime. Figure 9 has the representation of network lifetime. The red line denotes High Energy First (HEF) algorithm and the green line represents Dynamic Cluster Formation Method (DCFM). Thus, it is proved that the network lifetime is improved in Dynamic Cluster Formation Method (DCFM).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

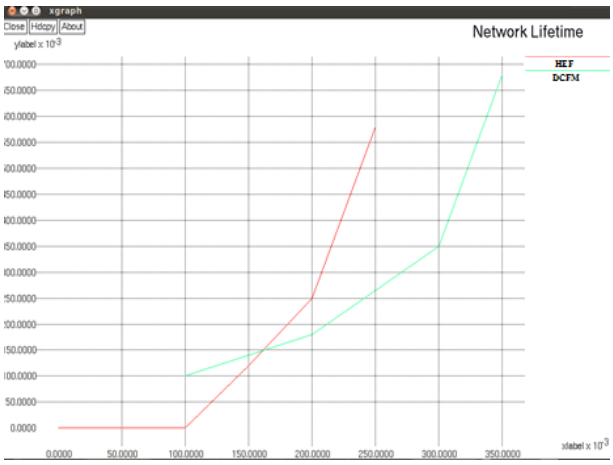


Figure 9. XGraph of Network Lifetime

- Energy of a node: Figure 10 represents the energy of a particular node. It is seen that the energy of the node is maximum at the initial state. Data transmission, data reception causes loss of energy in a node. It is seen that the node's energy drains eventually. At a particular threshold level, another node is made the cluster head. A cluster is formed only with the help of cluster head. The selection of cluster head is an important part. Thus, the selection of cluster head is dependent upon two major parameters. The residual energy and the delay time. The node with maximum residual energy and minimum delay time is made the cluster head. The delay time is calculated based on the node count. Node count is the number of broadcast message a node receives. With the node count, the delay time is calculated. From the calculated delay time, it is possible to determine the cluster head depending upon its residual energy.

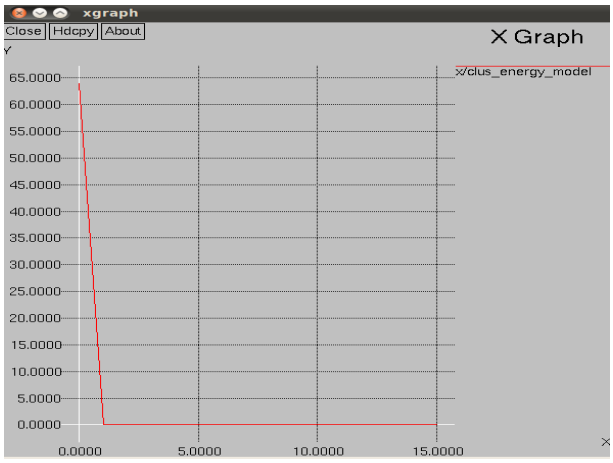


Figure 10. Energy of a Node

- Delay of a Node: Figure 11 denotes the delay of a node. The node with minimum delay and maximum residual energy is made the cluster head and it is used in Dynamic Cluster Formation Method (DCFM).

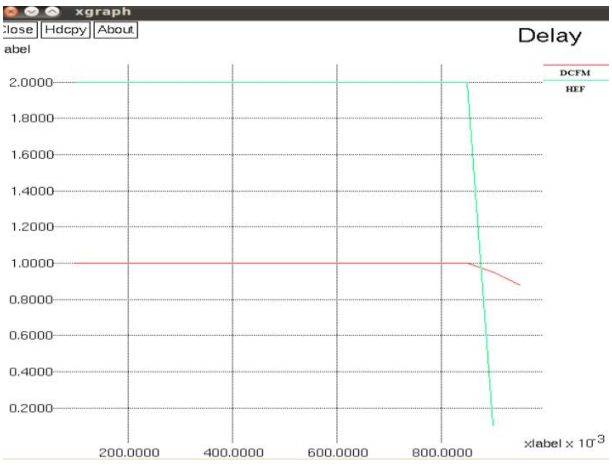


Figure 10. Delay of a Node

The residual energy and the delay time are important in evaluating the network performance. Thus the node with the minimum delay time and maximum residual energy is made the cluster head according to the threshold value. By doing this the energy is utilized efficiently thereby increasing the network performance.

VIII. CONCLUSION

On providing a trustworthy system behavior with a guaranteed hard network lifetime is a challenging task to safety-critical and highly-reliable WSN applications. For mission critical WSN applications, it is important to be aware of whether all sensors can meet their mandatory network lifetime requirements. In this project, we have addressed the issue of the predictability of collective timeliness for WSNs of interests. First, the Dynamic Cluster Formation Method (DCFM) is proven to be an optimal cluster head selection algorithm then, provide theoretical bounds on the feasibility test for the hard network lifetime. As there is an enhancement only in the network lifetime for now, there would be a greater chance of increasing the coverage and connectivity of the wireless sensor network (WSN) with a balanced energy consumption and an increased network lifetime.

REFERENCES

[1] Aurenhammer. F, "Voronoi Diagrams-A Survey of a Fundamental Geometric Data Structure," ACM Computing Surveys, vol. 23, pp. 345-405, 1991.

[2] Bandyopadhyay. S and Coyle. E.J, "An Energy Efficient Hier- archical Clustering Algorithm for Wireless Sensor Networks," Proc. IEEE INFOCOM 2003, Apr. 2003.

[3] Bertsekas. D.P, Nonlinear Programming, second ed. Belmont, Mass.: Athena Scientific, 1999. J. Bacon, D.M. Eyers, J. Singh, and P.R. Pietzuch, "Access Control in Publish/Subscribe Systems," Proc. Second ACM Int'l Conf. Distributed Event-Based Systems (DEBS),2008.

[4] Chen. H, C.-S. Wu, Y.-S. Chu, C.-C. Cheng, and L.-K. Tsai, "Energy Residue Aware (ERA) clustering algorithm for leach- based wireless sensor networks," in Second International Conference on Systems and Networks Communications (ICSNC 2007).

[5] Dongyao Jia, Huaihua Zhu, Shengxiong Zou and Po Hu., "Dynamic cluster head selection method for wireless sensor network," IEEE Sensor Journal, June 2015.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[6] Ghiasi et al. S, "Optimal Energy Aware Clustering in Sensor Networks," MDPI Sensors, vol. 2, no. 7, July 2002.

[7] Govindan. R, Hellerstein. J.M, Hong. W, "The Sensor Network as a Database," Technical Report 02-771, Computer Science Dept., Univ. of Southern California, Sept. 2002.

[8] Hansen. E, Neander. J, Nolin. M, and Björkman. M, "Energy-efficient cluster formation for large sensor networks using a minimum separation distance," in In The Fifth Annual Mediterranean Ad Hoc Networking Workshop, Lipari, Italy, June 2006.

[9] Hu Bin,Chen Wen. Analysis of Wisdom Agricultural Systems based on Wireless Sensor Network[J]. Sensors and actuators chemical,2013,(8):57.

[10] Intanagonwiwat. C, Estrin. D, Govindan. R, and Heidemann. J, "Impact of Network Density on Data Aggregation in Wireless Sensor Networking," Proc. 22nd IEEE Int'l Conf. Distributed Computing Systems, pp. 457-458, 2002.

[11] Javaid N, Waseem M, Khan Z A. ACH: Away Cluster Heads Scheme for Energy Efficient Clustering Protocols in WSNs[C]. 2013 Saudi International Electronics, Communications and Photonics Conference, Piscataway: IEEE.2013:364-7.

[12] Johnson. D, Maltz. D, and Hu. D, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)," IETF Internet draft, Feb. 2002.

[13] Kang.I and Poovendran. R, "Maximizing Static Network Lifetime of Wireless Broadcast Ad Hoc Networks," Proc. IEEE Int'l Conf. Comm. (ICC '03), 2003.

[14] Kaur. H and Baek. J, "A strategic deployment and cluster-header selection for wireless sensor networks," IEEE Transactions on Consumer Electronics, vol. 55, no. 4, November 2009.

[15] Liu H. Energy-balanced clustering algorithm for wireless sensor networks (EBCA) [J]. Electronic Measurement Technology, 2015.

[16] Mainwaring. A, Polastre. J, Szewczyk. J, Culler. D, and Anderson. J, "Wireless sensor networks for habitat monitoring," in Proc. WSNA, Atlanta, Georgia, Sept. 2002.

[17] Min. R, Bhardwaj. M,Cho et al.S. H, "Low-power wireless sensor networks," in VLSI Symp. Tech. Dig., Jan. 2001.

[18] O.B. Akan *et.al*, "Event-to-sink reliable transport in wireless sensor networks".

[19] Raghunathan. V et al., "Energy-Aware Wireless Microsensor Networks," IEEE Sig. Proc. Mag., vol. 1, no. 2, Mar. 2002.

[20] S. D. Muruganathan *et.al*, "A centralized energy-efficient routing protocol for wireless sensor network".

[21] Shah R C, Rabaey J M. Energy aware routing for low energy ad hoc sensor networks[C]. Prof.Of the 3rd IEEE Wireless Communications and Networking Conf.(WCNC),Orlando,2013:151-165.

[22] The Network Simulator – NS-2, http://www.isi.edu/nsnam/ns/

[23] Tyagi S, Kumar N. A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks[J]. Journal of network and computer applications: 2013,36(2):623-645.

[24] Wang Sheng-Shih, Chen Hung-Chang, Chen Ze-Ping " An Energy and Link Efficient Clustering Technique for Reliable Routing in Wireless Sensor Networks"(IEEE), 2012.

[25] Wang.W and Jantsch. A., "An algorithm for electing cluster heads based on maximum residual energy," in Proceeding of the 2006 international conference on Communications and Mobile Computing, July 03-06, 2006.

[26] Wei-Peng Chen et.al, "Dynamic clustering for acoustic target tracking in wireless sensor networks".

[27] Weifa Liang et.al, "Online data gathering for maximizing network lifetime in sensor networks".

[28] Yong Yuan *et.al*, "Virtual mimo-based cross-layer design for wireless sensor networks".

Context based Power Aware Multi-Effector Action optimized Reinforcement Learning

Mr. S. Senthil Kumar, Research Scholar. Dr. T. N. Ravi, Assistant Professor, Periyar EVR College

Abstract*— Multi-Effector Action Optimized Reinforcement Learning provides a configurable intruder detection system with dynamic security procedure switching schemes using one of the best Machine Learning (ML) procedures Reinforcement Learning (RL). An automated 'security threshold determining procedure' based on the active heterogeneous network circumstances is provided here to operate with Reinforcement Learning in the name of "Context based Power Aware Multi-Effector Action optimized Reinforcement Learning" (CPAMEA-RL). This procedure finalizes the security threshold values based on the context of the data. This value is important to choose an optimum security scheme which works on pre-calculated computational-power guidelines, so that the network security administration is provided with amended power utilization.

***Index Terms*— Reinforcement Learning (RL), Machine Learning (ML), Multi-Effector Action optimized Reinforcement Learning (MEA-RL), Context based Power Awareness, Security threshold determining, security based on computational-power guidelines.**

I. Introduction

Modern communication mostly carried out by a number of clusters of mixed type electronic network nodes. This heterogeneous network communication has a wide range of data and bandwidth utilization. Communication protocols and security policies among the cluster nodes are mostly diversified based on the nodes categories. Most of the nodes are battery operated or rechargeable at least and they equipped with beneficial mobility. This precariousness nature of nodes makes the clusters dynamic and causes the entire network into a less predictable entity. Providing the best security for this network without affecting its Quality of Service (QoS) is a challenging job. The QoS of any network is depending on the standard network parameters of Throughput, Communication Delays, Level of Security, and Power consumption.

Increasing the positive factors of QoS like throughput, security while decreasing the negative factors like jitter, latency, End-to-End delay and Power consumption is the vital aspiration while designing a raw network architecture. Intervening manually entire communication of this modern network pattern consumes more computational resources whereas the results are not up to the mark. This situation makes the manual security monitoring as a desolate task. Providing automatic security to this network improves the QoS because of the modern Machine Learning and Artificial Intelligence procedures. A substantially good Network Simulator like OPNET can be used to create a replica of the real world modern heterogeneous network along with existing and proposed security models the benchmark parameters like throughput, jitter, latency, end-to-end delay, security and power consumption can be measured. A number of simulations are performed with random network node placements and with random communications the benchmark results are measured and tabulated. These tabulated values are used to calculate the significance level of QoS improvement by using statistical calculations.

II. Related Works

A number of automated security policy selection schemes are contrived in the past decade. The major classification of these schemes based on Artificial Intelligence [1], Neural Network [2], Machine Learning [3] and Data mining technologies [4]. Some of the hybrid security policy schemes are running based on combining multiple technologies in simultaneous or sequential mode. Based on the standard network QoS parameters, upmost qualified technologies are selected into comparison.

The selected procedures are

1. Reinforcement Learning (RL)
2. Reinforcement Learning based on MDP (RL-MDP)
3. Multi-Effector Action Reinforcement Learning (MEA-RL)

1. Reinforcement Learning (RL):

Reinforcement Learning overcomes the disadvantages of its former procedures and performed well with dynamic independent data. RL combines both active and passive approach learning simultaneously. RL adopts the natural

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This **work** was supported in part by the U.S. Department of Commerce under Grant BS123456".

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

learning method of midbrain dopamine that learns by performing reward oriented prediction. Each knowledgebase entry of RL resembles the actual firing of a dopamine neuron. RL periodically updates its knowledgebase based on 'State-Action and Reward State-Action'. These knowledgebase updates are used to effectively train the system even with independent data. Thus Reinforcement Learning is recommended to introduce Artificial Intelligence (AI) based network security.

RL considers all 41 essential network security data factors from KDD-Cup dataset [5]. They are duration of the connection, type of the communication, service type, communication flag, Number of source data Bytes, Number of destination data Bytes Geographical location, Fragment Errors, Priority, Communication mode, Number of failed logins, Login flag, compromised network connections, Root access, Number of root access attempts, Number of file creations, Number of shell access, Number of files accessed, Number of outbound commands, Host Login flag, Guest Login flag, Login counts, Service Count, Service error count, connection error rate, service discard rate, guest-host service ratio, guest-host differential service rate, Service differential ratio, Number of destination hosts, Destination host service count, Destination host same service rate, Destination host different service rate, Destination host port match rate, Destination host server different host rate, Destination host server error rate, Destination host server service error rate, Destination host error rate, Destination host server recent error rate.

All these parameters are involved in calculating decision making factors for RL. Expected sum of immediate and long time rewards under the more suitable policy referred as Utility. It is calculated as

$$util(s_t, a) = E\left\{\text{Reward}(s_t, a) + \max policies \sum_{j=1}^{N-1} R_{t+j}\right\}$$

Where s_t refers the state at particular timestamp t, $\text{Reward}(s_t, a)$ refers immediate reward of executing action a in state
s_t, N refers number of steps taken by the agent in its lifetime. E{.} refers expectation over all possible combination of decisions.

Sometimes RL abides by taking reward oriented heuristic decisions makes the security system vulnerable to strategic long term attacks. In this criterion RL needs larger time consuming updates in its knowledgebase which makes the security system less responsible to the real-time data.

2. Reinforcement Learning with Markov's decision Process (RLMDP):

Knowledge base updates in RL are consuming more time against 'strategic attacks' and this problem is solved in RLMDP. Markov's Decision Process reduces many inutile heuristics movements performed in RL. Whenever there is an ambiguous decision or a decision with less support count occurs, RL took the default action expecting a reward whereas RLMDP applies MDP and filters the action if there is a less probability to get the reward. This nature of RLMPD makes it more stable against different attacks.

Markov Decision Processes (MDPs)[6][7] are operates on high dimensional state and action spaces represented as s and a respectively. To get the state s_t, action a_t and reward r_t at time t the state transition combinational probabilities and expected reward function is declared as $P(s_{t+1}|s_t, a_t)$ and $R(s_t, a_t)$. Stochastic and stationary polices declared by conditional distributions over actions $\pi^\theta(a; s)$ parameterized by θ. It is assigned that given policy π^θ the MDP is ergodic with stationary distribution d^θ. In RLMDP energy-based policies are considered which can be expressed as conditional joint distributions over actions a and a set of latent variables h

$$\pi^\theta(a, n; s) = \frac{1}{Z(s)} e\phi(s, a, h)\top_\theta \rightarrow (1)$$

where *φ(s,a,h)* are a pre-defined set of features and $\Sigma_{a,h} exp(e\phi(s, a, h)\top_\theta)$ is the normalizing partition function. The policy itself is then obtained by marginalizing out h. Latent type variables used to make policies based on energy and these policies classify composite non-linear and non-product relationship between actions and states inherent classification (1) is log linear with the features *φ(s,a,h)*. In a conditional restricted Boltzmann machine (RBM), the states *s*, actions a and latent variables h are all high dimensional binary vectors, and (1) is parameterized as

$$\pi^\theta(a, n; s) = \frac{1}{Z(s)} e^{s\top w_s h + a^\top w_a h + b_s\top s + b_h\top h + b_a\top a} \rightarrow (2)$$

where the parameters are matrices *Ws, Wa* and vectors *bs, ba, bh* of appropriate dimensionalities. Marginalizing out *h*, used to get a non-linearly parameterized policy

$$F^\theta(s, a) = -b_s\top s - b_a\top a - \Sigma_i \log\left(1 + e^{s^\top w_{si} + a^\top w_{ai} + b_{hi}}\right),$$
$$\pi^\theta(a; s) = \frac{1}{z(s)} e^{-F(s,a;\theta)} \rightarrow (3)$$

where *i* indices the latent variables, and w_{si}, w_{ai}, b_{hi} are parameters associated with latent variable h_i. The quantity $F(s, a; \theta)$ is the conserved energy.

The policy selection is constantly updated by SARSA, the state action pairs can be the nearest neighbor nodes. Here physical position of cluster information is used instead of Virtual Power Cluster (VPC) to reduce computational power. The error rate of SARSA can be computed as

$$\varepsilon(s^t, a^t) = [r^t + \gamma Q(s^{t+1}, t^{t+1})] - Q(s^t, a^t) \rightarrow (4)$$

In case the state-action function is determined by θ, then the update equation for new parameter is

$$\Delta\theta \propto \varepsilon(s^t, a^t)\nabla_\theta Q(s^t, a^t) \rightarrow (5)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The update process determines the security policy $M_x(R)$for the corresponding cluster P_x. The RL system was pre-trained to a beginning level with the optimal action

$$P(a|s) \approx \frac{e^{Q(s,a)/\tau}}{z} \rightarrow \quad (6)$$

where Z is a normalization factor, τ is a positive number represents the iteration. The RL convergence can be identified with the value of τ, if it is getting higher values, then it refers the RL training is under progress is with uniform improvement.

RLMDP operates with more power awareness than the other existing methods discussed here. The number of battery operated nodes is increasing in modern network. Therefore providing more security with less power consumption is important in modern network security systems. The concept of Virtual Power Clusters (VPC)[8][9] is used in RLMDP to facilitate a balanced action between power and security. The lack of parallelism and linear State Action - Reward State Action are main disadvantages of RLMDP and this affects the performance of RLMDP while dealing real-time data.

3. Multi-Effector Action Reinforcement Learning (MEA-RL):

Multi-Effector Action Reinforcement Learning (MEA-RL) is designed to use parallel sandboxing technique. Two sandboxed environments are set up in each cluster head to monitor all incoming connection requests. State action – Reward State action are performed in parallel based on the 41 network parameters of the incoming connection request and the reward attaining decision is updated in the knowledgebase. The decision which is miscarried the reward is inflicted in the knowledge base.

In MEA-RL, if $P(s)$ is a decision making policy with any of the mapping from states to actions, then the policy action quotient Q^P can be calculated as

$$Q^P(s_t, a_t) = E[r1+\gamma r_{i+1} + \gamma^2 r_{i+2} + \gamma^3 r_{i+3} \ \ldots | s_i, a_i, p]$$

Futures states can be calculated by performing recursive form as

$$Q^P(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}|s_t, a_t)\, Q^P(s_{t+1}, P(s_{t+1}))$$

Optimal value function along with associated policy can be calculated as

$$Q^*(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}|s_t, a_t) \max_{a_{t+1} \in A} Q^*(s_{t+1}, a_{t+1})$$

Consider the mean-estimate rule is $\mu_k(s_k)$, then error driven mean-estimation is calculated using

$$\mu_{k+1}(s_k) = \mu_k(s_k) + k_k \cdot \delta_k$$

where k_k is knowledge base update rate (learning rate) and it is calculated using

$$k_k = \frac{\sigma_k^2(s_k)}{\sigma_k^2(s_k) + \sigma_r^2(s_k)}$$

Prediction error is calculated using $\delta_k = \gamma_k - \mu_k$

III. Proposed Method & Implementation

Context based Power Aware Multi-Effector Action optimized Reinforcement Learning consists of 3 main concepts. CPAMEA-RL follows IPv6 protocols header format added with additional contents to incorporate the following concepts.

1. Data Sensitivity
2. Sensitivity Bits
3. Security Protocol Allocator

1. Data Sensitivity:

Modern Network data is an aggregation of multiple sensitive data. Each data has its own importance and sensitivity. Data Sensitivity module of CPAMEA-RL is designed to operate based on the data sensitivity regulation [10] recommended by Massachusetts Institute of Technology (MIT). As per the guidelines, data are classified in four sensitivity threshold. The highest security index data are Credit Card and Bank Account details, Social Security Numbers, Personal Medical Data, Military related documents and confidential data of Research Organizations. This kind of data should be kept confidential and should be handled with proper security authentications.

The second sensitivity category is high confidential index data like financial information, information disclosed by non-disclosure agreements, management information and contract details. These data are containing a highest security request tag in general, where there are two possibilities of security services subsists. If the host has enough power to process these data with high security authentication, then these data are treated like high security index data and security is gained by the highest authentication procedures. Another case, if the host is not having sufficient power to process these data with highest security protocols, second grade security protocols are followed to conserve power. In this case both power saving and highest security are not guaranteed but either one is assured.

The third sensitivity category is low confidential index data like social media forwarded data, public chat information, details of a shared or public library and discussion forums. These data are processed with low power consuming security procedures. Some amount of power saving is assured while handling these information in CPAMEA-RL. The fourth sensitivity category is a secure-free type data like public entertainment broadcasting data, streaming entertainment data and open libraries. These data are meant to

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

be prepared to reach almost every node without any authentication requirement from the sender side. The host has the rights to block these data individually by marking as spam or permitting these data in which power is used only for communication and not for any security crypto procedure. CPAMEA-RL is designed to handle all these four kinds of data sensitivity in a desirable mode.

2. Sensitivity Bits:

Sensitivity bits are used to mark the sensitivity of the data. Two bits are used here since there are four sensitivity classifications in CPAMEA-RL.

00 – is used for security free communication data
01 – Low security index data
10 – High security index data
11 – Highest security index data

These two bits are added as the extension header with the standard 40 Bytes IPv6 data header [11]. Structure of IPv6 standard header is given in Table 1.

S.No	Bits	Description
1	4	Version
2	8	Traffic Class
3	20	Flow Label
4	16	Payload Length
5	8	Next Header
6	8	Hop Limit
7	128	Source Address
8	128	Destination Address

[Table 1]

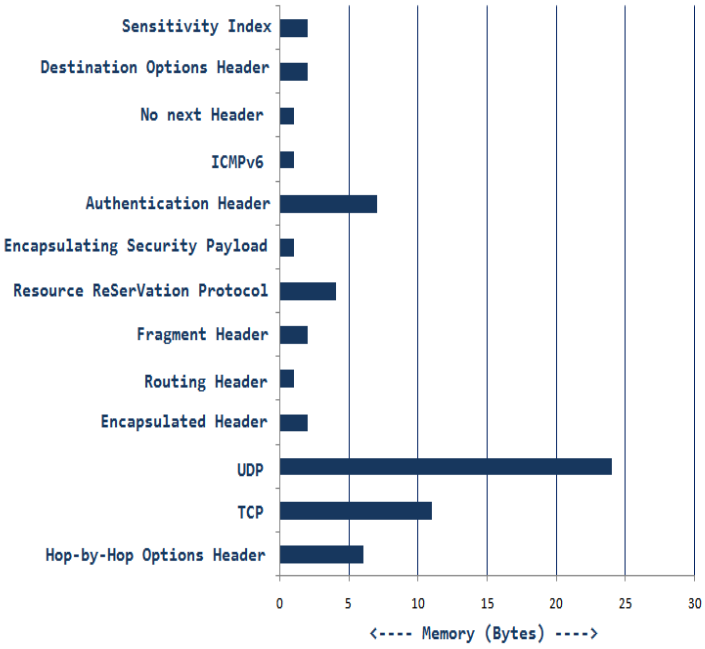
The standard extension header of IPv6 is given in table 2

Extension Header	Next Header Value	Description
Hop-by-Hop Options header	0	read by all devices in transit network
Routing header	43	Contains methods to support making routing decision
Fragment header	44	Datagram fragmentation parameters
Destination Options header	50	read by destination devices
Authentication header	51	Authenticity Information
Security Payload header encapsulation	60	Destination Options Header

[Table 2]

Sensitivity bits are added after destination options header in bit positions 61 and 62. IPv6 has similar information that assigns the security payload header encapsulation in destination options header. The difference is destination options header of the IPv6 protocol is assigned by the sender and will be processed only by the receiver. The intermediate nodes and cluster heads are not processing the security header where as the sensitivity bits are designed to process by the cluster heads. Cluster heads are authorized to allocate security resources based on the sensitivity bits' values. Since sensitivity bits are added as the last header in IPv6 protocol's header sequence, the value 59 is assigned as the next header field that refers nothing follows the sensitivity bits.

CPAMEA-RL packet header is illustrated in picture 1.



[Figure 1: CPAMEA-RL Packet Header]

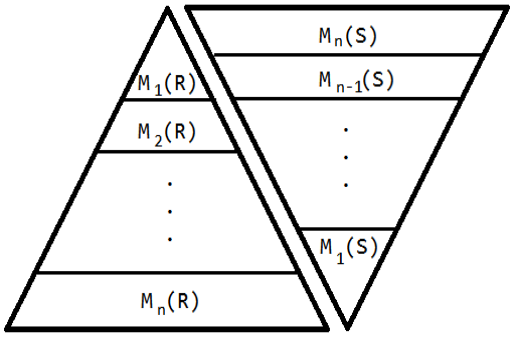
3. Security Protocol Allocator:

The actual use of Context based Power Aware Multi-Effector Action optimized Reinforcement Learning is utilized in this module. Real data sensitivities are in many layers whereas they categorized into four major types (with two reserved bits). So each major security category is consists of multiple security level layer. Allocating a suitable security protocol for each network communication that occurs in same sensitivity category with different security layer.

The Reinforcement Learning System is equipped with Memory mapping of Security Protocols. The memory map is created using two parameters they are Resource Consumption and Security Strength shown in Figure 2. In figure 1, the first triangle refers the resource consumption

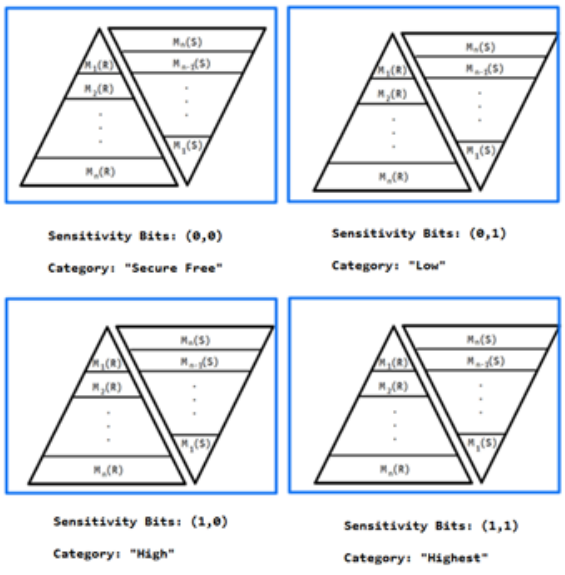
https://sites.google.com/site/ijcsis/
ISSN 1947-5500

of the security protocols from $M_1(R), M_2(R) \ldots M_n(R)$ and they are arranged in ascending order based on the resource consumption that is $M_1(R)$ consumes very less resource than any other method and $M_n(R)$ consumes the highest resource than the other methods taken in to consideration. The second triangle refers the security strengths $M_1(S)$ to $M_n(S)$ of the security protocols from $M_1(R)$ to $M_n(R)$ respectively in ascending order. That is $M_n(R)$ provides the highest security and $M_1(R)$ provides the least security while comparing with the other participating methods.



[Figure 2: Memory mapping of security protocols]

In CPAMEA-RL, the four major sensitivity categories are allocated with corresponding Memory maps with different security protocols. Each major sensitivity category uses RL to find out a desirable security protocol for the sub-security layer to involve in the communication shown in Figure 3.



[Figure 3: CPAMEA-RL Security Protocol Allocator]

The process of selecting the sensitivity category is deterministic because of the debut of sensitivity bits and selection of security protocol in a sensitivity category is non-deterministic, thus RL is applied here to solve the problem. The security protocol aggregation contains seven cryptographic procedures Rivest-Shamir-Adleman Algorithm (RSA), Data Encryption Standard (DES), Triple Data Encryption Standard (3DES), Advanced Encryption Standard (AES), Elliptic Curve Cryptography (ECC), Blowfish and International Data Encryption Algorithm (IDEA). These cryptographic procedures are configured to use different size keys based on the requirement.

AES, DES and 3DES are used predominately in Low security Sensitive Category. Low Power utilization and less computational work are involved in this sensitivity category. RSA and IDEA are used mostly in high security sensitivity category. Moderate power utilization with adequate security is achieved by using these procedures. Blowfish and ECC are used with comparatively large keys in the highest sensitivity security segment. Power is compromised here but security is the prime concern of this high secure zone.

KDD-Cup dataset is used to train CPAMEA-RL in a similar way which is used to train RL. The difference is, in CPAMEA-RL, Multi-Effector Action Optimization reduces the training time. The dataset contains 3,925,650 attacks and 972,781 normal records (4,898,431 transactions in total) are adequate to make the CPAMEA-RL.

In CPAMEA-RL, if $P(s)$ is a decision making policy with any of the mapping from states to actions and δ is the sensitivity index, then the policy action quotient Q^P can be calculated as

$$Q^P(s_t, a_t) = E[r1+\gamma r_{i+1} + \gamma^2 r_{i+2} + \gamma^3 r_{i+3} \ldots | s_i, a_i, d, p]$$

Futures states can be calculated by performing recursive form as

$$Q^P(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}, \delta_{t+1} | s_t, a_t)\, Q^P(s_{t+1}, \delta_{t+1}, P(s_{t+1}))$$

Optimal value function along with associated policy can be calculated as

$$Q^*(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}, \delta_{t+1} | s_t, a_t) \max_{a_{t+1} \in A} Q^*(s_{t+1}, \delta_{t+1}, a_{t+1})$$

Consider the mean-estimate rule is $\mu_k(s_k)$, then error driven mean-estimation is calculated using

$$\mu_{k+1}(s_k) = \mu_k(s_k) + k_k \cdot \delta_k$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

where k_k is knowledge base update rate (learning rate) and it is calculated using

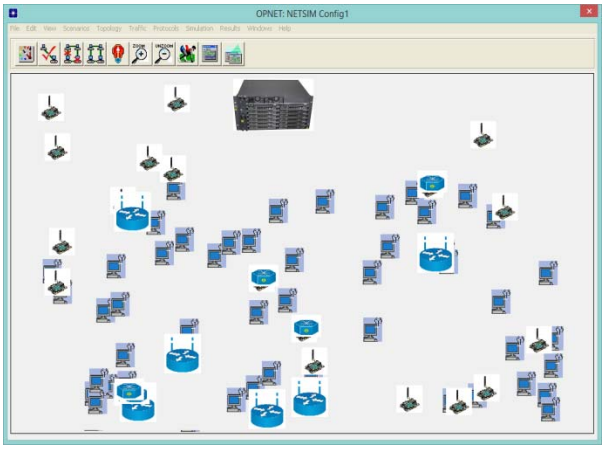
$$k_k = \frac{\sigma_k^2(s_k)}{\sigma_k^2(s_k) + \sigma_r^2(s_k)}$$

Prediction error is calculated using $\delta_k = \gamma_k - \mu_k$

IV. Performance Analysis

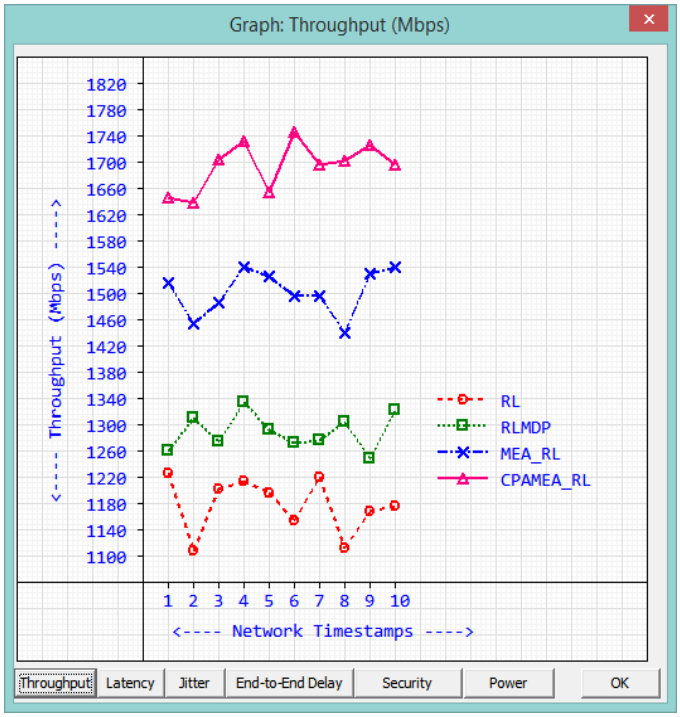
Performance of MEA-RL along with existing methods are measured by calculating the standard network QoS parameters throughput, latency, jitter, end-to-end delay, security level and average power consumption. Data are evenly distributed into four security labels ($\{\delta_0, \delta_1, \delta_2, \delta_3\}$). The data are classified into four types as Control Data (D_c), Text data (D_T), Voice data (D_v) and Multimedia data (D_m). All these types are evenly distributed in typical heterogeneous network traffic. Ten equal time stamps are selected from the simulation process. Active Reinforcement Learning, Reinforcement Learning, and Reinforcement Learning with MDP are taken as the participants in the simulation to compare with the proposed Multi-Effectors Action Reinforcement Learning. An User Interface and script wrapper to OPNET Modeler[12][13] is designed with Visual C++ programming language of Visual Studio 2013 Integrated Development Environment(IDE). NETSIMCAP – a Network Simulation and Capture Software Development Kit is used to interface Visual C++ with OPNET network simulator. Centralized Server with six Wi-Fi routers and 50 heterogeneous nodes are placed using random placement scheme[14] of OPNET to create a hybrid heterogeneous network[15][16] environment. OPNET creates required virtual network environment then process the exact network scenario while measuring the parameters instead of performing calculations. By this fact the results from OPNET are more realistic than any other calculation based results.

Example CPAMEA-RL network scenario for security enhancement: Security Level δ_x is classified as $\forall\delta_i$, i from 0 to n, $\delta_0 \leq \delta_x \leq \delta_{n-1}$ where δ_0 represents least security and δ_{n-1} represents most security. Available power resource σ_x is between least power index σ_0 to most power index σ_n are assorted in VPCs. When data with security index δ_{n-2} arrives to a VPC with power index σ_n, then δ_{n-2} will be elevated to the next security level of $\delta_{n-1}$where security is ensured. Example CPAMEA-RL network scenario for Energy efficiency [17][18]: When data with security index $\delta_1$arrives to a VPC with power index σ_0, then security index $\delta_1$will be imposed to a lesser security level δ_0. In some network transactions, this security step down activity may assailable but compromised security will not be a problem because the context is pre-assigned to a low sensitivity type by the sensitivity bits.



[Fig.5 OPNET Network structure]

Figure 5 shows the placement of heterogeneous nodes and wireless network distribution hotspots in OPNET

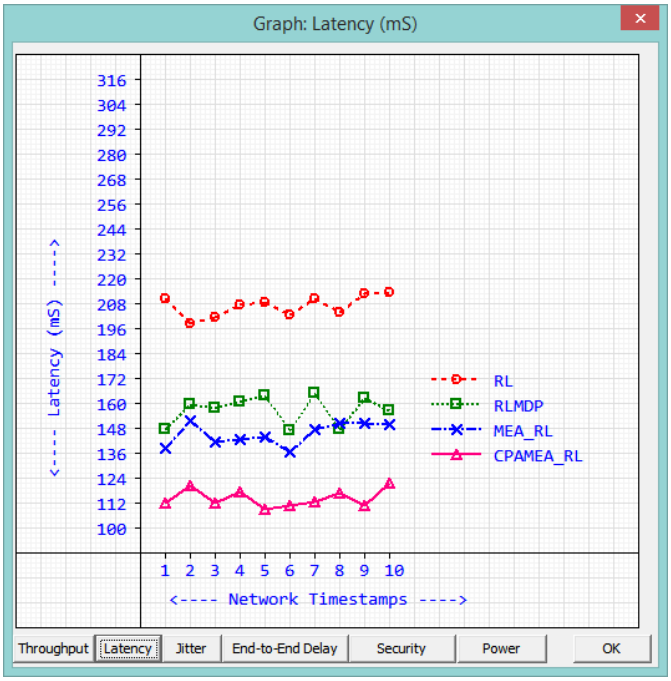


[Fig.6 Throughput Mbps]

Throughput of RL, RLMDP, MEA-RL and CPAMEA-RL are shown in Figure 6. RL achieved throughput from 1108 Mbps to 1226 Mbps based on the random locations of the nodes.

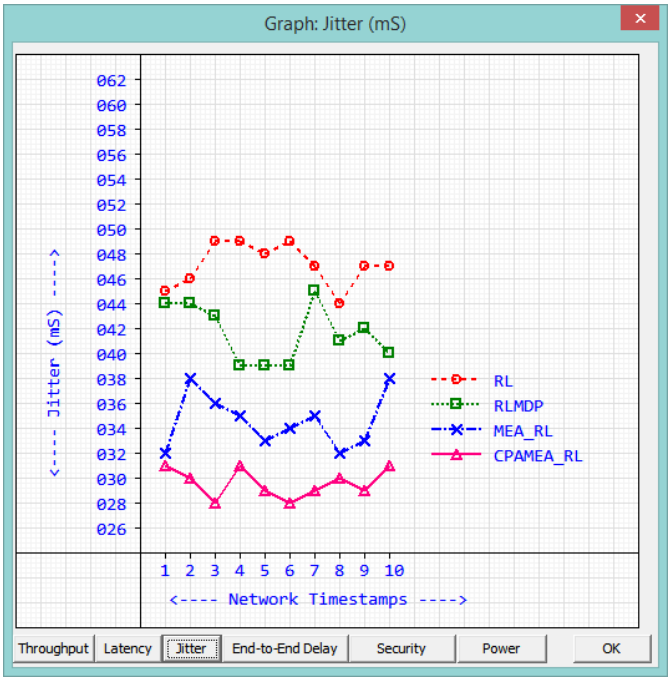
RLMDP achieved a little better than RL, i.e. from 1249 Mbps to 1334 Mbps. MEA-RL got the throughput range of 1440 Mbps to 1541 Mbps which is higher than RL and RLMDP. Whereas proposed CPAMEA-RL achieved the highest throughput of 1639 to 1746 Mbps range which is higher than all other methods taken into comparison – shown in Figure 6.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



[Fig.7 Latency (mS)]

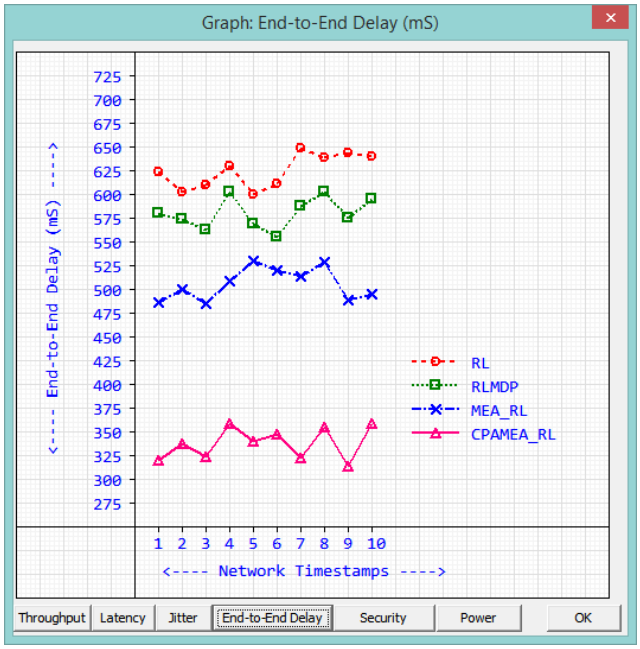
Latency is a delay in nodes response for a network transaction. This latency is inversely proportional to QoS of a network. To maintain a better QoS latency has to be kept to the bare minimum negligible level. CPAMEA-RL reduces the latency to the minimum value of 109 mS whereas other existing methods took 137 to 214 mS. Latency comparison chart of existing and proposed methods is given in figure 7.



[Fig 8. Jitter (mS)]

The time difference in packet inter-arrival time to their destination is called as jitter. Jitter is a natural delay in packet based network communication. In general, TCP and IP protocols are dealing with the jitter impact on communication. To achieve higher QoS, jitter should be kept to the minimum negligible level.

The lowest value of 28mS is achieved by the CPAMEA-RL implies the higher performance than the other methods involved. Comparison graph is shown in Figure 8.

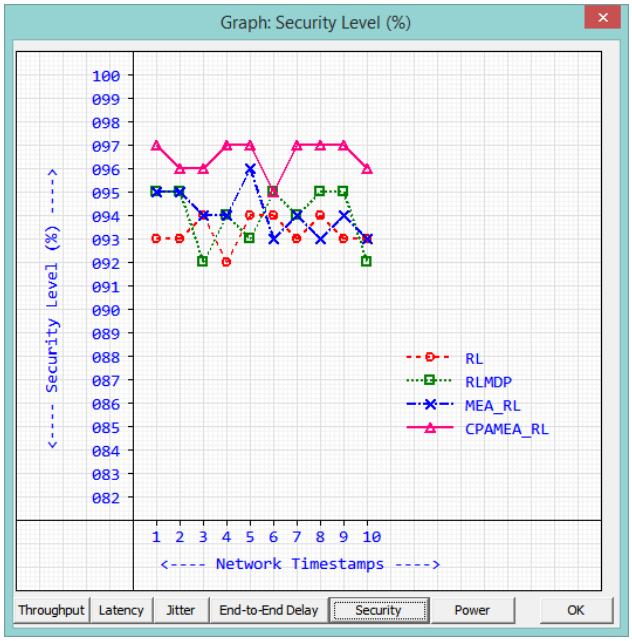


[Fig 9. End-to-End Delay]

Average travelling time taken by a data packet from source to destination is called as End-to-End Delay. It includes delays caused by route discovery process and the data packet transmission queue. Dropped packets are not considered while calculating end-to-end delay and all successfully delivered packets are included in the end-to-end delay calculation.

The measured End-to-End delay of CPAMEA-RL method is shown in Figure 9. CPAMEA-RL gets the minimum end-to-end delay of the range from 314mS to 359mS.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



[Fig.10 Security Level (%)]

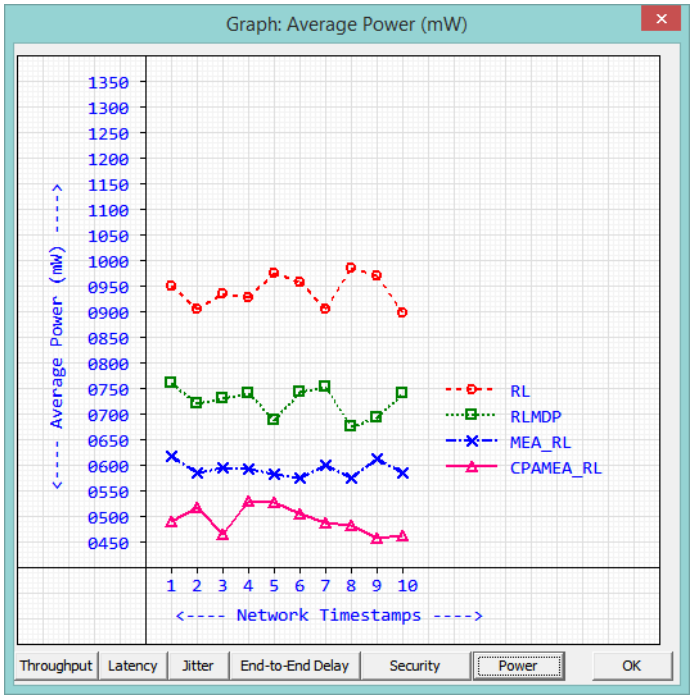
Security is the vital criteria involved in modern networks with shared and distributed infrastructures. The higher security level refers the higher quality of the network architecture. The highest security value of 97% is achieved by CPAMEA-RL is shown in Figure 10. Even though RL and RLMDP are getting closer security levels with the security level of proposed CPAMEA-RL, higher category average is achieved by CPAMEA-RL.

The security strength is measured by OPNET simulator's internal mechanism which consists of all standard attacks like Brute force attack, Dictionary attack, Wormhole attack, Sinkhole attack and etc. The measured security for MEA-RL and CPAMEA-RL are given in Table 3. CPAMEA-RL achieved 96.5% whereas MEA-RL achieved 94.1%. The improvement in security of 2.4% is a significant improvement when the security scores are above 90%.

Time stamp	MEA-RL(%)	CPAMEA-RL(%)
1	95	97
2	95	96
3	94	96
4	94	97
5	96	97
6	93	95
7	94	97
8	93	97
9	94	97
10	93	96
Average	94.1	96.5

[Table 3]

The prime target of proposed method is to provide uncompromising QoS with highest security and lowest power consumption. Based on the OPNET simulation measurements, CPAMEA-RL is used lesser power range from 459mW to 530mW. Average power consumption of existing methods with proposed methods are compared in Figure 11.



[Fig.11 Average Power (mS)]

Average power consumption is measured for MEA-RL and CPAMEA-RL is given in table 4. CPAMEA-RL used 99.5mW lesser than the MEA-RL on average. Measured power for MEA-RL and CPAMEA-RL are given in table 4.

S.No	MEA-RL(mW)	CPAMEA-RL (mW)
1	619	492
2	587	518
3	596	467
4	594	530
5	583	528
6	575	507
7	602	488
8	576	484
9	614	459
10	585	463
Average	593.1	493.6

[Table 4]

V. Conclusion

In this paper, the Reinforcement Learning method is endued with innovative Context based Power Aware Multi-Effctor Actions. Based on the observed results in a typical heterogeneous network simulation environment, CPAMEA-RL secures highest QoS indicants. The crown part of a security system is to provide highest security with lowest power consumption which is achieved in CPAMEA-RL. Since

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

CPAMEA-RL is equipped with the cutting-edge technologies, it is ready to be used in the process of constructing robust heterogeneous network environments.

REFERENCES

[1] Daniel Grzonka, Agnieszka Jakobik, Joanna Kolodziej, Sabri Pllana. Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security- Future Generation Computer Systems Elsevier June 2017

[2] Min-Joo Kang, Je-Won Kang. Intrusion Detection System Using Deep Neural Network for In-Vehicle Network Security. PLOS One Tenth Aniversary June 2016

[3] Anna L.Buczak, Erhan Guven. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials October 2015

[4] IH Witten, E Frank, MA Hall, CJ Pal. Data Mining: Practical Machine learning Tools and Techniques. Google Books 2016

[5] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A Detailed Analysis of the KDD CUP 99 Data Set – IEEE 2009

[6] Deniz Gunduz, Kostas Stamatiou, Nicolo Michelusi, Michele Zorzi. Designing intelligent energy harvesting communication systems. IEEE Communications Magazine - IEEE January 2014

[7] Mugen Peng, Shi Yan, Kecheng Zhang, Chonggang Wang. Fog-computing-based radio access networks: issues and challenges. IEEE Network Volume: 30, Issue: 4 - IEEE August 2016

[8] Ado Adamou Abba Ari, Blaise Omer Yenke, Nabila Labraoui, Damakoa, Abdelhak Gueroui. A power efficient cluster-based routing algorithm for wireless sensor networks: Honeybees swarm intelligence based approach. Journal of Network and Computer Applications - Elsevier April 2016

[9] Zhongyuan Zhao, Mugen Peng, Zhiguo Ding, Wenbo Wang, H.Vincent Poor. Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks. IEEE Journal on Selected Areas in Communications Volume: 34, Issue: 5 - IEEE March 2016

[10] Juan C. Asenjo. Data Masking, Encryption, and their Effect on Classification Performance: Trade-offs Between Data Security and Utility. Nova Southeastern University http://nsuworks.nova.edu/gscis_etd/1010. June 2017

[11] P. Thubert, Ed, C. Bormann, L. Toutain, R. Cragie. IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Routing Header. RFC Editor RFC 8138 https://www.rfc-editor.org. April 2017

[12] M. Fazeli, H. Vaziri. Assessment of Throughput Performance Under OPNET Modeler Simulation Tools in Mobile Ad Hoc Networks (MANETs). Computational Intelligence, Communication Systems and Networks, 2011 Third International Conference. ISBN: 978-1-4577-0975-3. IEEE 2011

[13] Kuinam J. Kim, Nikolai Joukov. Communications-based technology for smart grid test bed using OPNET simulations. Information Science and Applications (ICISA). Springer 2016

[14] Mingyue Ji, Giuseppe Caire, Andreas F.Molisch. Wireless Device-to-Device Caching Networks: Basic Principles and System Performance. IEEE Journal on Selected Areas in Communications - IEEE July 2015

[15] Hamnah Munir, Syed Ali Hassan, Haris Pervaiz. Energy Efficient Resource Allocation in 5G Hybrid Heterogeneous Networks: A Game Theoretic Approach. Vehicular Technology Conference - IEEE March 2017

[16] E. Bodanese. A brief introduction to heterogeneous networks (HetNets) and its challenges. Communication Technology and Application (ICCTA 2011), IET International Conference. IET- 2011

[17] Elhadj Benkhelifa, Tom Welsh, Loai Tawalbeh, Yaser Jararweh, Anas Basalamah. Energy Optimisation for Mobile Device Power Consumption: A Survey and a Unified View of Modelling for a Comprehensive Network Simulation. Mobile Networks and Applications - Springer August 2016

[18] Uchenna Odih, Panos Bakalis, Predrag Rapajic. Investigating the Impact of Traffic Type on the Performance of on Demand Routing Protocols and Power Consumption in MANET. Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015 - Springer January 2016

COMPARATIVE ANALYSIS OF CONVERGENCE TIMES BETWEEN OSPF, EIGRP, IS-IS AND BGP ROUTING PROTOCOLS IN A NETWORK

Eunice Domfeh Asabere [1], Joseph Kobina Panford [2], James Ben Hayfron-Acquah [3]

[1] Final Year M Sc. I.T, Dept. Of Computer Science, KNUST, Kumasi, Ghana, easabere@ssnit.org.gh
[2] Dept. Of Computer Science, KNUST, Kumasi, Ghana, jpanford@yahoo.com
[3] Dept. Of Computer Science, KNUST, Kumasi, Ghana, jbha@yahoo.com

---

Abstract - *Convergence time is a key factor in determining performance of routing protocols and routing protocol is one of the significant factor in determining the quality of IP communication. Convergence time is therefore very essential to a network and networks that converge faster are considered to be very reliable. The research was carried out to compare the convergence of four routing protocols namely OSPF, EIGRP, IS-IS and BGP. Network scenarios were created and a simulation was performed using Graphic Network Simulator (GNS3) to measure the convergence times of the protocols separately. Results indicated that EIGRP had the fastest convergence time in both link failure and topology change scenarios. This will help network administrators in their choice of protocols.*

***Key Words*: Convergence time, Protocol, Network, Routing, OSPF, BGP, IS-IS, EIGRP**

1. INTRODUCTION

1.1 Background of Study

Data packets traveling through the network typically traverse multiple routers and thus multiple physical links interconnecting them. Whenever there is a link failure or change in topology, routing protocols try to provide an alternative path towards the destination. It is therefore crucial that the routing protocol quickly detects such a link failure or topology change. With the increasing use of networks, any unnecessary loss of connectivity can hardly be tolerated and has to be kept as short as possible. This brings up the issue of convergence time and network is believed to have converged when the routing tables on all routers within the network are complete and correct. Routing protocols play a major role in the delivery of packets from source to destination addresses. In the study, four routing protocols namely Open Shortest Path First (OSPF), Border Gateway Protocol (BGP), Intermediate system to Intermediate system (IS-IS) and Enhanced Interior Gateway Routing Protocol (EIGRP) were compared to determine their convergence time in a given network topology.

1.2 Statement of the Problem

One of the most important characteristics of routing protocols is the convergence time The convergence time determines how fast the routers adapt their routing tables to topological changes. Among OSPF, EIGRP, IS-IS and BGP, a proof-based advice for selecting the one with the best convergence time is aimed at.

1.3 Research Objectives

The main objectives of this research are:

- To determine the convergence time for OSPF, BGP, IS-IS and EIGRP in a particular network topology.
- To compare the performance of OSPF, BGP, IS-IS and EIGRP.

1.4 Research Questions

The following are the research questions that were posed in order to accomplish the objectives.

- What is the convergence time for OSPF, BGP, IS-IS and EIGRP in a network Topologies?
- Which of the four routing protocols has the fastest convergence time in the topology used.

1.5 Significance of the Study

This study will be significant in providing an in depth understanding of the four routing protocols OSPF, BGP, IS-IS and EIGRP and determining the convergence time of these protocols in a network. The study will also compare the convergence times of these routing protocols and come out with the best one.

2. LITERATURE REVIEW

Convergence can be defined in many ways but in the context of computer networks, a network is said to have converged when all routers in a network have the same topological information about their network they find their selves in. With the help of routing protocols, routers collect topological information [1]. Convergence is a critical property in routing especially dynamic routing. There are about three forms of routing namely static, default and dynamic [2]. A network topology is said to have converged "when routing tables on all routers within the network are complete and correct" [3]. Convergence addresses the manner in which networks recover from problems and network changes. Modern networks anticipate problems by providing alternate, redundant or standby paths.

Convergence time is the time that is required for the routers in a network to learn about routes in a given network. This time is important because it helps administrators of a network to determine in the event that a network downtime occurs due to a failed link between routers or any damage to one router the amount of time it will take for that network to recover and begin to function as a normal network.

Deng et al. [4] performed analysis of RIP and OSPF and EIGRP using OPNET which is a simulator widely used for networking related analysis. In their research, they analysed the performance of these protocols based on their convergence activity, convergence duration and traffic sent

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(bytes/sec) to compare the difference in their performance. From their research, they found out that the convergence of EIGRP was faster than the others regardless of the network topology.

Panford *et al*. [5] also analyzed Convergence times between RIP and EIGRP routing protocols in a network using packet tracer which allows network behaviour experimentation and also helps in answering what-if scenarios. In their research, they observed that EIGRP had the fastest convergence time. Their experiments also showed that, regardless of the topology, the convergence time remains the same whether for RIP or EIGRP. Another interesting observation made with EIGRP was that as the number of routers increases, the time for convergence were almost the same.

3. METHODOLOGY

The method for this research was a simulation of scenarios. To help with this simulation, Graphical Network Simulator (GNS3) was employed and the network diagram for the simulation scenarios is illustrated in Fig.1. GNS3 was chosen because it has a user- friendly Graphical User Interface (GUI) and also enables users to configure a network component in a virtual machine that runs the OS same as the original network component.

4. ANALYSIS

The measurements results were placed into three main categories. The first category, based on Fig. 2 consist of measurements of convergence times of protocols with link failure closer to the source of the traffic as shown in Table 1. The second category as derived from Fig 3. is made up of measurements of convergence times of protocols with link failure closer to the destination of the traffic as illustrated in Table 2 and the last category was convergence time measurements under topology change as shown in Table 3. Fig. 4 and Fig. 5 show the network diagram as additional routers are added to the original network diagram.

4.1 Results of Routing protocols with Failure closer to the source of the traffic

Table -1: Convergence time measurement for Protocols with Link Failure closer to the source of the traffic.

Test	OSPF	EIGRP	IS-IS	BGP
1	8.346	6.078	7.799	21.975
2	8.637	5.985	8.340	23.169
3	8.494	5.938	8.368	19.331
4	8.879	5.951	8.185	15.275
5	8.162	5.469	8.590	15.743
6	9.601	6.039	9.095	19.874
7	7.788	5.491	8.532	25.646
8	8.592	5.169	8.042	28.507
9	8.836	7.298	8.051	14.664
10	6.187	6.204	7.441	28.640

4.2 Results of Routing protocols with Failure closer to the destination of the traffic

Table -2: Convergence time measurement for Protocols with Link Failure closer to the source of the traffic.

Test	OSPF	EIGRP	IS-IS	BGP
1	7.993	4.220	8.091	13.553
2	8.490	3.969	8.485	29.517
3	7.797	4.079	8.344	31.220
4	9.078	4.007	8.438	29.993
5	9.005	5.968	8.905	16.394
6	8.673	5.972	8.297	18.859
7	8.938	4.298	7.801	27.502
8	8.841	6.001	8.438	20.646
9	7.735	5.875	8.660	17.641
10	7.704	5.969	7.438	24.204

4.3 Results of Topology Change

Table -3: Convergence time measurement for Protocols with Topology Change

	OSPF	EIGRP	IS-IS	BGP
1	12.138	3.027	8.044	18.227
2	11.117	3.031	12.148	20.258
3	10.913	3.468	11.082	18.229
4	11.530	3.477	9.275	19.541
5	10.026	3.102	9.196	19.205
6	10.993	3.198	10.084	18.714
7	11.362	3.144	9.143	21.095
8	11.122	3.112	8.012	19.521
9	11.212	3.099	8.050	18.008
10	10.410	3.005	7.048	18.035

On the average, it took OSPF network 8.352s to converge, EIGRP network 5.962s to converge, IS-IS network 8.244s to converge and BGP network 21.282s to converge for link failure closer to source of the traffic. For link failure close to the destination of traffic the average convergence times were 8.375s for OSPF, 5.036s for EIGRP, 8.290s for IS-IS and 22.953ms for BGP.

It took an average time of 11.082s for OSPF to converge, 3.166s for EIGRP to converge, 9.208s for IS-IS to converge and 19.083s for BGP to converge for change in topology.

5. CONCLUSION AND RECOMMNDATION

From the simulation results the EIGRP give the best performance. EIGRP generate the least traffic and thus it will consume the least bandwidth, leaving enough bandwidth for transmission of data. EIGRP also has the best performance in the case of topology changes and when there is a broken Ethernet connection.

In conclusion, the simulations confirmed that EIGRP was the best choice for all scenarios implemented as it has a fast convergence, while also efficiently utilizing bandwidth.

IS-IS was the second choice as far as convergence time was concerned and then OSFP came next. BGP performed poorly

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and is therefore not suitable for large networks. It can therefore be stated based on the results achieved that there is a significant difference in the performance of the protocols as far as convergence time is concerned.

REFERENCES

[1] Shah A, Waqas J. Rana, "Performance Analysis of RIP and OSPF in Network Using OPNET", International Journal of Computer Science Issues, Issue 6, No 2, November 2013.

[2] Lammle, Todd, "CCNA Cisco Certified Network Associate study guide, sixth edition". Indianapolis, Ind.: Wiley. (2007).

[3] Todorovic Ivana, Sepanovic Stevan, "Measurements of convergence time for RIP And EIGRP Protocols", Scripta Scientiarum Naturalium, volume 2, 2011.

[4] Deng Justice, Wu Siheng, Sun Kenny, "Comparison of RIP, OSPF and EIGRP Routing Protocols based on OPNET" Final project. Spring, 2014.

[5] Panford, J. K., Riverson K. & Boansi O. K (2015). Comparative analysis of convergence times between rip, and eigrp routing protocols in a network. Research Journali's; Journal of Computer Science., pg 1-5.

APPENDIX

The following are network topologies used in the experiments.

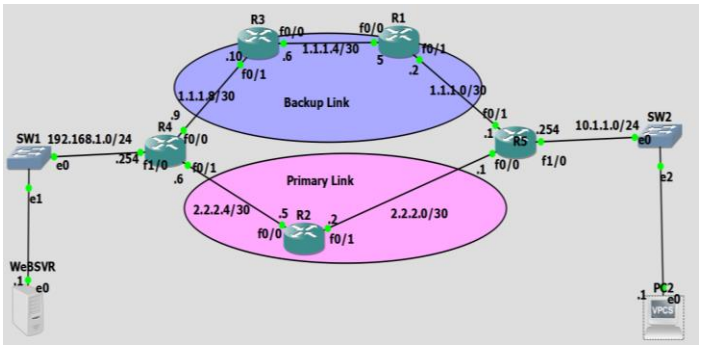


Fig -1: Network diagram for simulation scenarios

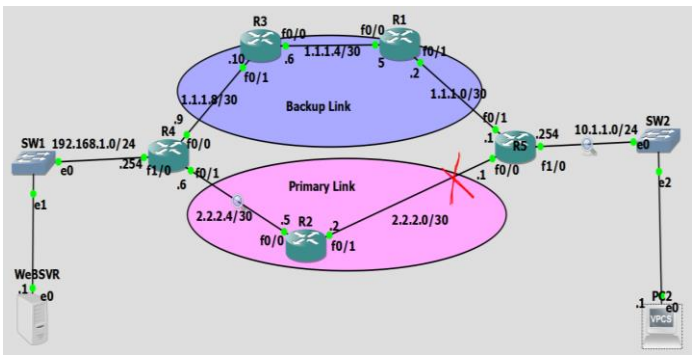


Fig -2: Network diagram for link failure closer to the source of traffic

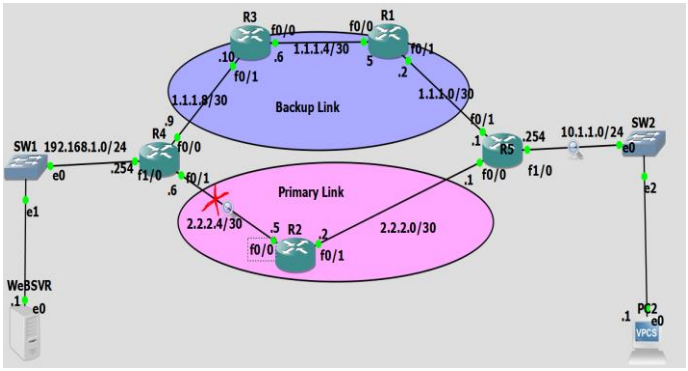


Fig -3: Network diagram for link failure closer to the destination of traffic

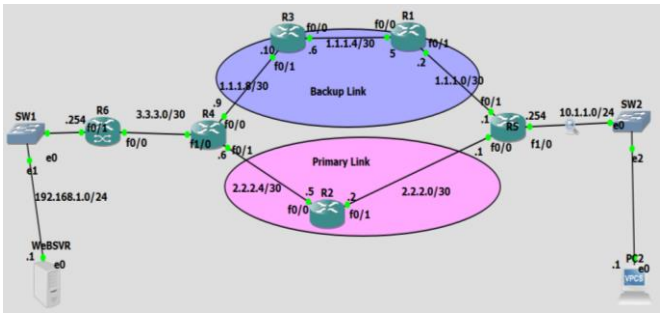


Fig -4: Network diagram with one additional router

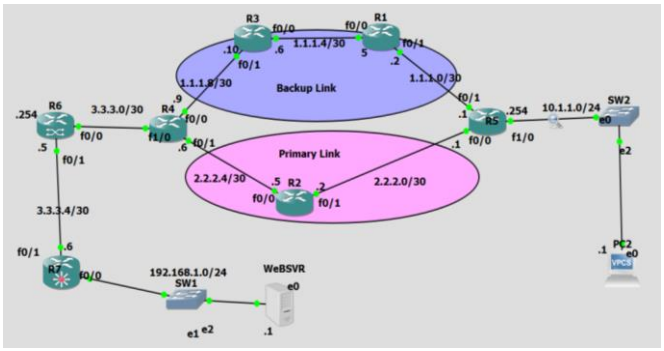


Fig -5: Network Diagram with two additional routers

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ARABIC SPELLING CHECKER ALGORITHM FOR SPEECH RECOGNITION

Huda F. Al-shahad

Department of Computer Science
University of Kerbala
Karbala, Iraq
huda_msc2006@yahoo.com

ABSTRAT

The Automatic Speech Recognition is defining as the process of convert a speech wave into text by using a computer. Speech recognition is the easiest way manipulate with the computer application especially to the people that have no arms. This paper proposes an Arabic word and popular language (Iraqi language) error correction method and algorithm for speech recognition system. The proposed algorithm is split the input content (that is input as a speech wave and convert it to text by speech recognition system) into a few word-tokens that are submitted as search questions to the system. The system offer to replace the error word by the suggested correction using n-gram features and save the writing words in a text file that the user will choose the path of it. Future research can improve upon the proposed system so much so that it can be take many correction algorithms and make difference between them.

Keywords: *Speech Recognition; Arabic Error Correction; popular language; Token; n-gram.*

I. INTRODUCTION

Speech technology is presently broadly utilized as a part of the field of discourse chronicling, for example, PodCastle [1]. In these frameworks, the words are perused by client or to recover the fitting sections utilizing watchwords, a low word-blunder rate (WER) is hardly must require, so the model must the most suitable words between the hopefuls assumed by an automatic speech model. Be that as it may, if many words in model are false, it might be chosen independent of what is the dialect display. This problem need solve, a few distinguish language models [3, 4, 5] have been proposed to re-rank the N-best sentences after large-vocabulary, continuous speech recognition.

The use of N-grams trained from speech recognition results including false words and it given transcription. This paper describes a method that receive the text from speech system after convert it and correct the error words by suggest the correction words that make the user have the flexibility to choose any one or replace the error word with the true word. After that the correction words will save in a text file. Many propelled discourse acknowledgment frameworks utilize trainable dialect models that can be advanced for a specific (speaker-free) and in addition for a particular sub-dialect use. This enhancement is important to accomplish a respectable level of acknowledgment precision; be that as it may, it may not ensure reliably high-exactness execution because of the constrained abilities of the basic dialect display, generally 2-or 3-gram HMMs. The method in this paper is to take the Arabic and popular language (Iraqi language) word (one or more) that the speaker says and make the correctness on it if it error. Unlike other approaches (e.g., Bassil& Alwani, 2012) that use the suggestion technology Bing's spelling to recognize error and correct the words that input by automatic speech recognition that recognized output text[5]. The other paper (Nishizaki & Sekiguchi, 2006) describes an error correction method of continuous speech recognition using WEB documents for spoken documents indexing [6]. Fusayasu, i Tanaka, Takiguchi and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Ariki in 2015 focus on their research on a word-error correction system for continuous speech recognition using confusion networks[7].
The structure of this paper is as follows. In Section 2, discuss Arabic challenges. In Sections 3 the error detection of text, the methodology is discussed in section 4, error correction is described in Section 5 and in 6 Computation algorithm is describe, and the experimental results are shown in section 7. Finally, the conclusion is view in Section 8.

II. CHALLENGES OF ARABIC SPEECH

Arabic speech recognition faces many challenges. One of these is the vowels of Arabic word are short which are may be ignore in text. Another one is Arabic language has many tones where each word is pronounced in a different way.
Arabic many-sided quality is appearing by the expansive number of affixes (prefixes, infixes, and suffixes) that can be added to the three shape design. Farghaly and Shaalan in 2009 gave an investigation of Arabic dialect difficulties and answers for it[8]. Lamel et al. in 2009 introduced many number of difficulties for Arabic discourse acknowledgment, for example, very large lexical variety[9].

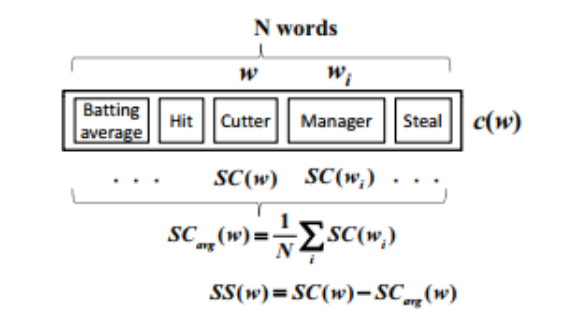


Figure (2) Semantic score

Similarity SC(wi) between the context c(w) and the number of word wi in the context is computed by latent semantic analysis (LSA) [10].

III. LATENT SEMANTIC ANALYSIS

" Latent Semantic Analysis (LSA) is a hypothesis and technique for separating and speaking to the importance of words. Significance is evaluated utilizing factual calculations connected to a huge corpus of content. The corpus encapsulates an arrangement of common limitations that to a great extent decide the semantic likeness of words and sets of words. These requirements can be understood utilizing direct variable based math strategies, specifically, Singular Value Decomposition."4 LSA is a numerical and factual approach, guaranteeing that semantic data can be gotten from a word-record co-event network and words and reports can be spoken to as focuses in a (high-dimensional) Euclidean space.

Dimensionality diminishment is a basic piece of this inference. LSA depends on the Vector Space Model (VSM), a mathematical portrayal of content archives generally utilized as a part of data recovery. The vector space of an accumulation of writings is built by speaking to each report as a vector containing the frequencies of the words or terms the record is made out of as components. By and large, these archive vectors signify a term-by-report framework speaking to the full content accumulation. Relatedness of records can be gotten from those vectors, e.g. by figuring the edge between archive vectors by methods for a cosine measure. In any case, this numerical portrayal of content information does not illuminate commonplace issues of working with dialect. From one perspective there are morphological issues for the correct recognizable proof of terms and the way that not all terms in content are of equivalent significance. This can be settled by highlight determination methods (stemming, stop word expulsion, collocations, equivalent word records, space vocabulary, grammatical form taggers, and data pick up) and weighting plans (TF-IDF, Log-Entropy). Solitary Value Decomposition (SVD) is utilized as a rank bringing strategy down to truncate the first vector space to uncover the hidden or 'inactive' semantic structure in the example

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

of word utilization to characterize archives in a gathering. This truncation permits managing common dialect issues like synonymy as various words communicating a similar thought should be near each other in the diminished k-dimensional vector space. SVD will break down the first term-by-report network into orthogonal components that speak to the two terms and records:

$$\mathbf{A = U \Sigma VT} \qquad \mathbf{(1)}$$

With A the original term-by-document matrix, Σ a diagonal matrix with the square roots of singular values of A.AT and AT .A ($\sigma 1^2 > \sigma 2^2 > \ldots > \sigma n^2$), and U and V containing left and right singular vectors [].

We will generate the document-word matrix by using tf-idf as shown in the following equation:

$$\mathbf{TFIDFi,j = (Ni,j / N^*,j) * log(D / Di)} \quad \mathbf{(2)}$$

After that the document is factored using singular value decomposition (SVD) as follows;

$$\mathbf{W = USV^T} \qquad \mathbf{(3)}$$

Using the row vector ui of the matrix U and the row vector vj of the matrix V , the similarity sim(ri, cj) between the document cj and the word ri is computed as follows:

$$\mathbf{sim(ri, cj)} = \frac{\mathbf{uiSvT\ j}}{|\boldsymbol{uiS}^{\frac{1}{2}}\ ||\boldsymbol{uiS}^{\frac{1}{2}}\ |} \qquad \mathbf{(4)}$$

IV. COMPUTATION ALGORITHM

The semantic aim of any word is defined to be high if the meaning of the selected word is similar to the meaning of the words around the underlining word. The semantic result of the word w is computed as follows:

(1) c(w) is represent the context of content word w that framed as the gathering of the substance words around w including a similar word, as appeared in Figure(2).

(2) The similarity SC(wi) between the context c(w) and the word wi in the context is computed, where i represent the number of word.

(3) The average similarity SC(wi) is computed as SCavg(w).

(4) Normalized similarity SS(w) is computed the difference between SC(w) and SCavg(w) as shown in equation below;

$$SS(w) = SC(w) - SCavg(w)$$

V. ERRORS DETECTION

Error detection problem may be solved by two techniques :*N-gram* analysis and *dictionary lookup*. Error correction method consists of checking to know if the input string is valid or not.

In this paper the n–gram will use. N-gram method is characterized as a strategy to discover erroneous words in content. Rather than looking at each time every whole word in content to the appropriate lexicon, n-grams will control the checking. The checking is finished by utilizing a matrix with an n-dimensional size where the frequencies of real n-gram are put away. On the off chance that a nonexistent or unusual n-gram is discovered then the word is flagged as an incorrect spelling word, else not.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A n-gram is an arrangement of successive characters brought from a content with a length of n as is set to.
When n = 1 character then the term that utilized is a **Unigram**,
When n = 2 characters then the term that utilized is a **Bigram**,
When n = 3 characters then the term that utilized is **Trigram**.

VI. MRTHODOLOGY

The proposed method in this paper shows the process of the error detection model using N-gram and LSA information to show the cost of error correction in system as shown in figure (1). The first step, speech data are recognized and the recognition results are output as a token. Second, each word is marked as false or true. After the recognition errors, the system will suggest the correction words that make the user choose which correct one need, and then the words will save in a text file in any drive preferred.
In this paper, as said above, word-mistake amendment can be accomplished in the perplexity set by choosing the word with the most astounding estimation of the accompanying straight discriminant work. We utilize the best probability words in the disarray arrange if the perplexity set has no third probability word, it is supplemented with the second one. Also, on the off chance that it has no second probability word, it is supplemented with the first. After the learning procedure is done, acknowledgment blunders are rectified utilizing the calculation beneath:

(1) Receive the text as voice, so we will convert the voice to test.

(2) Make tokenize and concentrate the best probability words from the perplexity organize and detect recognition.

(3) Using the error detection model, “N-gram”.

(4) Apply the LSA algorithm to find the suitable word.

(5) Select the best likelihood word in the confusion set if the word identified as correct data does not exist.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

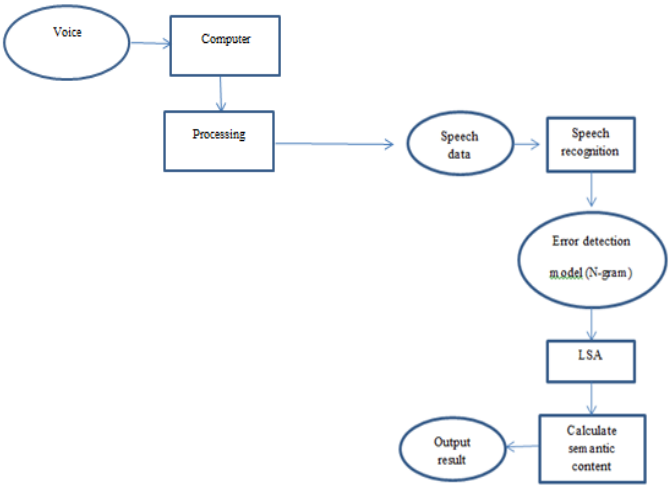


Figure (1) proposed method

VII. ERROR CORRECTION

The proposed error correction algorithm includes several steps that must execute in order to detect the error and then correct it. The algorithm take-off by divide the recognized output transcript into many tokens T={ t1…tn }, each composed of n words, ti={w0,w1,w2,w3,w4,..wn} where ti is represent the special token and wj is a single word in that token. Then, every ti is sent to check the validation of it using n-gram to check the ranking and show the suggestion correct words ci. If the word is valid, at that point token ti must not contain a specific incorrectly spelled word; and thus, ti is supplanted by ci. At last, after all tokens get approved, all the first right tokens O={ t1…tk }, plus the corrected ones C={ c1…cp } will concatenate with each other, to make a new text with fewer error represented formally as V={ v1…vk+p }. In this paper, we use the characteristic N-gram. To put it plainly, we utilize it to distinguish acknowledgment blunders. This sort of discriminative dialect demonstrate can be prepared by fusing the discourse acknowledgment comes about what's more, the comparing right interpretation. Discriminative dialect models, for example can distinguish unnatural N-grams and adjust the false word to fit the characteristic N-gram.

VIII. EXPERIMENT AND RESULTS

In the experiments, speech recognition was performed in to two different languages: English and Arabic (clear Arabic language and Iraq language).The proposed calculation was executed utilizing MS C# 4.0 under the MS .NET Framework 4.0 and the MS Visual Studio2012.The speech recognition will enter the words that the user want to print as in figure (3) and then the system will lexical the text into many tokens, take the token to check if the word error or not, if the word is error many suggestion will appear to replace it by the word as in figure (4). The correct output will save in word text, the figure(5) shows the Iraqi word that the user may be entered when talk with family in another place for example or when two friends talk to each other by computer or through any social media program.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

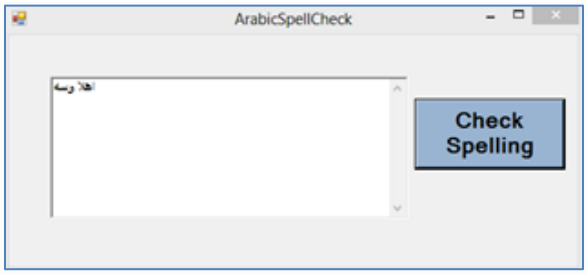


Figure (3) Speech Input

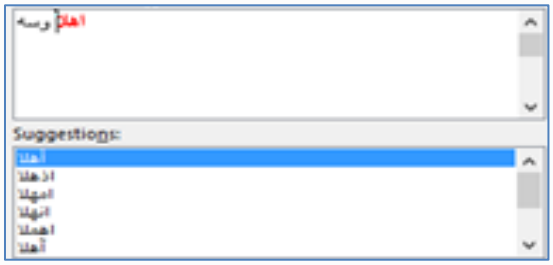
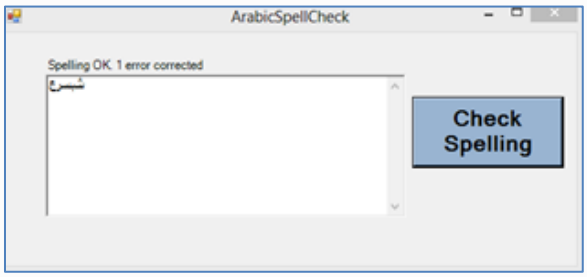
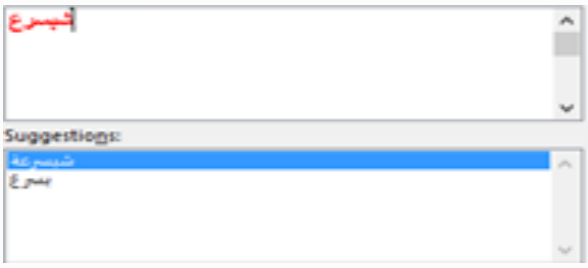


Figure (4) Correction suggestion



figure(5-A) Iraqi words



figure(5-B) Iraqi word correction

When we use the LSA algorithm the count matrix that output from the Arabic text is as follow:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
[[ 1.  2.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 2.  1.  0.  0.  1.  0.  0.  0.  0.]
 [ 0.  1.  1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 0.  1.  0.  0.  2.  2.  2.  0.  2.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 0.  0.  0.  0.  0.  1.  1.  1.  1.]]
```

figure(6) count matrix

The reason SVD is valuable, is that it finds a diminished dimensional portrayal of our lattice that underscores the most grounded connections and discards the commotion. As it were, it makes the most ideal recreation of the framework with the minimum conceivable data. To do this, it tosses out commotion, which does not help, and underlines solid examples and patterns, which do help as shown in follow:

```
Here are the singular values
[  5.47054849e+00   3.24054292e+00   1.54975161e+00
   8.69865786e-01   7.43977135e-01   6.15341902e-17
   4.27433567e-35]
```

figure(7) SVD vector

After that the similarity will use to check the validity of the word in the document, the figure (8-a) show the part of columns of the matrix and figure (8,b) show the part of raws of the matrix.

```
[[ 0.07868054 -0.61603335  0.50014023]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.1064358  -0.6735659  -0.46037367]
 [ 0.03566079 -0.23712335  0.57850259]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.73209861 -0.11018105 -0.30750493]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.30657657  0.15175944  0.21449164]]
```

figure(8-a) column of matrix

```
[[  5.32948634e-02   1.88565420e-01   6.51868729e-03  -2.77555756e-17
    2.87107044e-01   5.40366902e-01   5.40366902e-01   5.60412855e-02
    5.40366902e-01]
 [ -6.05813657e-01  -6.95234426e-01  -7.31739566e-02   1.38777878e-17
   -2.75857479e-01   1.48317161e-01   1.48317161e-01   4.68314850e-02
    1.48317161e-01]
 [ -2.71402911e-01   5.23248020e-01   3.73287298e-01  -0.00000000e+00
   -6.93907026e-01   6.46682950e-02   6.46682950e-02   1.38403886e-01
    6.46682950e-02]]
```

figure(8-b) raws of matrix

IX. CONCLUSION

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this paper, I have proposed an automatic speech recognition Arabic and Iraqi language error correction by using n-gram algorithm. The proposed two-step first, speech data are recognized and split the text into tokens , second, each word is labeled as false or true and recognition errors, the system will suggest the correction words, correction method can efficiently use the n-gram method.

REFERENCES

[1] M. Goto, J. Ogata, and K. Eto, "Podcastle: A web 2.0 approach to speech recognition research," in Proc. Interspeech2007, 2007, pp. 2397–2400.

[2] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in Proc. ACL, 2004, pp. 47–54.

[3] T. Oba, T. Hori, and A. Nakamura, "A study of efficient discriminative word sequences for reranking of recognition results based on n-gram counts," in Proc. Interspeech2007, 2007, pp. 1753– 1756.

[4] A. Kobayashi, T. Oku, S. Homma, S. Sato, T. Imai, and T. Takagi, "Discriminative rescoring based on minimization of word errors for transcribing broadcast news," in Proc. ISCA, 2008, pp. 1574– 1577.

[5] Y. Bassil and M.Alwani, "Podcastle: Post-Editing Error Correction Algorithm For Speech Recognition using Bing Spelling Suggestion," in (IJACSA) International Journal of Advanced Computer Science and Applications2012, pp.95-101.

[6] H. Nishizaki and Y. Sekiguchi, "Word Error Correction of Continuous Speech Recognition Using WEB Documents for Spoken Document Indexing"2006, ICCPOL pp. 213–221.

[7] Y. Fusayasu, K.Tanaka, T. Takiguchi and Y. Ariki," Word-Error Correction of Continuous Speech Recognition Based on Normalized Relevance Distance",in Conference on Artificial Intelligence (IJCAI 2015), pp.1257-1262.

[8] A.Farghaly and Kh.Shaalan," Arabic Natural Language Processing: Challenges and Solutions ", 2009, ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4.

[9] L. Lamel, A. Messaoudi et al ," Automatic speech-to-text transcription in Arabic ",2009, ACM Transactions on Computational Logic, pp. 1–17

[10] T. Landauer, P. W. Foltz, D. Laham, "Introduction to Latent Semantic Analysis", in Discourse Processes, 1988, pp. 259–284.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Detection of physiological changes in women during muscle activity using electromyography and other techniques

Rossana Rodríguez Montero[a], David Asael Gutiérrez Hernández[a], Miguel Mora González[b], Víctor Manuel Zamudio Rodríguez[a], Juan Martín Carpio Valadez[a], Manuel Ornelas Rodríguez[a], Claudia Margarita Lara Rendón[a], Miguel Salvador Gómez Díaz[a], Jesús Ricardo Sevilla Escoboza[b], and Víctor Porfirio Vera Ávila[b]

a. Tecnológico Nacional de México. Instituto Tecnológico de León. León, Guanajuato, México.
b. Universidad de Guadalajara. Centro Universitario de los Lagos. Lagos de Moreno, Jalisco, México.

***Abstract*— Bio-signals are important to know what is going on with our body. Especially the muscular activity is related with physiological changes in a woman, for example with their menstrual cycle. Besides of this, it is required to evaluate muscular activity over this changes, this can be done with electromyography (EMG) and the entropy, which allow the comparison of the obtained signals to measure those physiological changes.**

In previous works, muscular fatigue has been evaluated with EMG; nevertheless it has not been going into deep with chemical changes that are produced into the body, in a natural way which can alter the obtained bio signal coming from the muscle.

We developed a digital portable electromyograph to get electromyography samples. By means of it, the women´s bio-signals were studied, for those who were under an exercise routine and also for those who were not.

While visualizing the behavior of the electromyography obtained from the muscle, we perceived the singularity that the bio-signal for women of both group, while being on a menstrual cycle were similar.

Thus, it was implemented the entropy on the signals to justify the results obtained on the electromyography and the personal test applied. As a result, we proved that these signals are really showing one of the physiological changes in a woman.

***Index Terms*— Electromyography (EMG); Entropy; Membrane potential; Muscle activity.**

I. INTRODUCTION

In the human body, the biological signals can come from many physical phenomes. For the case of this work, the bio-signal is taken from the biceps muscle area; but to be able to process and obtain conclusions from the muscle activity, first it has to be converted in signals of electric character [1]. To achieve the obtainment of that signal, electromyography is the selected technique to be applied.

The EMG is a biomedical signal that measures the electric current generated in the muscles during its contraction and represents the muscular activity. One of the most popular techniques for the acquisition of this kind of signals is the superficial electromyography, that is commonly used for many researches, being this a noninvasive technique that uses electrodes that are put on the skin environment for taking the differential of bio-potential created by the variations of current in the muscle cells [2,3] which can be useful for quantitative technique for evaluation and registration of the electric activity produced by muscles [1,4].The electromyography signals contains relevant information that may be used for patron detection of a signal [5], the progress of the muscular fatigue [6,7], among other systems or applications type[3].

We can say that this signals give a time serial of biological kind; on which the time in a signal point, can mean more than a

This work was supported in part by the Conacyt, and the Instituto Tecnologico de Leon's laboratory "LACIT".

Rossana Rodríguez Montero is with the Instituto Tecnológico de León, León, Guanajuato, México (e-mail: m105060381@itleon.edu.mx).

David Asael Gutierrez Hernandez, is with Department of postgrade and Intelligent environments department, Instituto Tecnológico de León, León, Guanajuato, Mexico (e-mail: david.gutierrez@itleon.edu.mx).

Miguel Mora González is with Department of Exact Sciences and Technology, Centro Universitario de los Lagos, Universidad de Guadalajara. Lagos de Moreno, Jalisco, México. (e-mail: mmora@culagos.udg.mx).

Víctor Manuel Zamudio Rodríguez is with Intelligent environments department, Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: vic.zamudio@ieee.org).

Juan Martín Carpio Valadez is with Intelligent Systems Department, Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: juanmartin.carpio@itleon.edu.mx).

Manuel Ornelas Rodríguez is with Intelligent Systems Department, Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: manuel.ornelas@itleon.edu.mx)

Claudia Margarita Lara Rendón is with Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: claus.l.r@icloud.com)

Miguel Salvador Gómez Díaz is with Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: miguel_gomez@itleon.edu.mx)

Jesús Ricardo Sevilla Escoboza is with Department of Exact Sciences and Technology, Centro Universitario de los Lagos, Universidad de Guadalajara. Lagos de Moreno, Jalisco, México. (e-mail: jesus.sescoboza@academicos.udg.mx).

Victor Porfirio Vera Avila is with Centro Universitario de los Lagos, Universidad de Guadalajara. Lagos de Moreno, Jalisco, México. (e-mail: viktorvera.182@gmail.com).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

simple analysis [8] in the muscular activity made.

The biological temporary serial data not necessarily can be evaluated with common methods used from time series analysis, like the techniques of autocorrelation and frequency domain [9].

With the previous explanation, the objective of this work is to identify changes or physiological states through the computational exploration of biological signals, obtained during muscular activity and by means of electromyography and entropy technique to evaluate this signal.

II. Theoretical Description

For the development of this work some theoretical concepts must be considered. There are the following:

A. *The membrane potential*

Neurons are the basic functional units of the nervous system, and they generate electrical signals called action potentials, which allow them to quickly transmit information over long distances.

The different classes of neurons that found in the human nervous system, can be divided into three classes: sensory neurons, motor neurons, and interneurons; where these have three basic functions, these are to receive signals (or information), integrate incoming signals (to determine whether or not the information should be passed along) and communicate signals to target cells (other neurons or muscles or glands).

That conjunction of neurons, when your brain decides to move a muscle, motor cortex neurons travel through the spinal cord to synapse with "lower motor neurons." These motor neurons on moment to make synapse with the muscle form a "motor unit", where a motor unit is composed of an individual motor neuron and many muscle fibers it innervates. A muscle fiber is a very special cell type that can change its shape thanks to the actin / myosin chains that travel in it [10].

An individual motor neuron can synapse with many muscle fibers. In general, a large muscle such as the biceps has motor neurons that innervate thousands of muscle fibers while other muscles, such as those in the eye, which require a lot of precision, have motor neurons that innervate less than ten muscle fibers [11].

When a motor neuron triggers an action potential, this potential generates a release of acetylcholine (fig. 1-2) at the synapse between the neuron and the muscle (this synapse is also known as Neuromuscular Junction). Acetylcholine causes a change in the electrical potential of the muscle. When this electric potential reaches a threshold, an action potential is generated in the muscle fiber this action potential propagates through the muscle membrane, causing the voltage-dependent calcium channels to open, which begins the cellular cascade that finally generates muscle contraction.

When you contract a muscle, it is because many muscle fibers are firing action potentials and changing their shape (fig. 2).

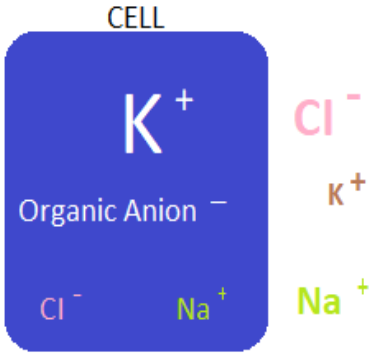


Fig. 1. Diagram of concentration and change of anions and cations, sodium (Na^+), potassium (K^+), chloride (Cl^-). (Image modified from "The sodium-potassium exchange pump," by Blausen staff (CC BY 3.0)).

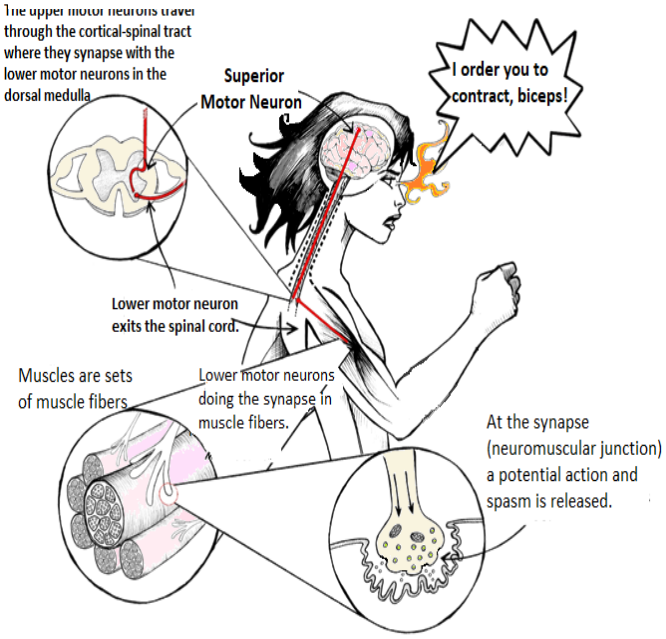


Fig. 2. Diagram of electrochemical process for the generation of muscle movement. Own elaboration

B. *Entropy*

According to [12] entropy is the degree of disorder that a system has and can be considered as a measurement standard. Entropy can be considered as a measure of uncertainty, so that the information needed in any process can be narrowed, reduced or eliminated uncertainty.

Entropy generation clarifies energy losses in a system evidently in many energy-related applications. Bejan [13] originally formulated the analysis of entropy generation.

1) *Permutation Entropy*

Permutation Entropy (PE) was introduced as a complexity parameter for time series based on comparison of neighboring values; the advantages are its simplicity, extremely fast calculation and robustness [14]. That kind of entropy is an appropriate complexity measure for chaotic time series, in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

particular in the presence of dynamical and observational noise. In contrast with all known complexity parameters, a small noise does not essentially change the complexity of a chaotic signal. Permutation entropies can be calculated for arbitrary real–world time series. As the article [15] says.

The article [16] says, that the algorithm to compute the PE can be divided into four basic steps, as the article:

1. Fragment the continuous EEG signal into segments containing *m* samples (*m* is called the embedding dimension); for a given embedding dimension *m* = 3 there will be *m*! possible permutations called motifs, so in this case six different motifs are obtained.
2. Identify each motif as belonging to one of the six different categories.
3. Obtain the probability of occurrence of each motif in the signal (pi) by counting the number of motifs of each of the six different categories.
4. Apply the standard Shannon uncertainty formula to calculate the PE of the resultant normalized probability distribution of the motifs (Eq. 1).

$$PE = \frac{-\sum(p_i \cdot \ln(p_i))}{\ln(number\_of\_motifs)} \quad (1)$$

C. Fourier transform

The Fourier transform (FT) has been approached from the formulation of the discrete signal, closer to its use in computable methods and algorithms, with its practical side of tool creation and applications in the treated field. The Fourier transform [17] represents a useful tool to extract the information contained in a signal on the domain frequency. FT is provided by its integral [18], that this provides a frequency function. That function is complex: its module is the spectral amplitude and the square from the amplitude is the density of spectral power (DSP) [19]. This spectral density is, the Fourier transform of the autocorrelation. The spectra term is used for the amplitude and for the power density represented in front of the frequency [19].

The DSP can be determined by some methods. The most used are the entropy method and the square of the Fourier transform. That is the reason why we used this two methods.

D. Hilbert Huang transform

The article [20] says, that Hilbert-Huang transform (HHT) is NASA's designated name for the combination of the empirical mode decomposition (EMD) and the Hilbert spectral analysis (HSA). It is an adaptive data analysis method designed specifically for analyzing data from nonlinear and nonstationary processes. The key part of the HHT is the EMD method with which any complicated data set can be decomposed into a finite and often small number of components, called intrinsic mode functions (IMF).

HHT is an empirical approach, and has been tested and validated exhaustively but only empirically. In almost all the cases studied, HHT gives results much sharper than any of the traditional analysis methods in time-frequency-energy representation [20]. Additionally, it reveals true physical meanings in many of the data examined. The Hilbert–Huang Transform (HHT) is a new time–frequency analysis method [21]. The main difference between the HHT and all other methods is that the elementary wavelet function is derived from the signal itself and is adaptive. The main feature of the HHT is the Empirical Mode Decomposition (EMD), which is capable of extracting all the oscillatory modes present in a signal. Each extracted mode is referred to as an Intrinsic Mode Function (IMF), which has a unique local characteristic [22, 23]. After the Hilbert transform on each IMF has been performed, the time–frequency distribution of the signal energy is obtained, which is referred to as the Hilbert spectrum.

III. Experimental Setup

For this experiment we used a digital electromyograph, developed in the laboratory of LACIT of the Technological Institute of Leon that serves as our system of acquisition of the muscular signal. This instrument has three cables that receive the signal, where an electrode is connected to each cable, these electrodes are placed on the skin of the test subject (figure 3); where two of them are in the middle and lower biceps (toward the elbow, calling these as positive and negative) and the third electrode is placed in the elbow area, as shown in figure 4; the reading obtained is saved in an extension file ".txt", then the information is analyzed on a computer. All of the above is the acquisition technique; for this work four samples were taken, one per week; seven women participate in the experiment, where they separate into 2 groups, those who exercise and those who do not exercise, their average age is 24.8 years. They responded to a brief questionnaire with their physical characteristics, performed exercise, and in the physiological phase that was (in menstrual cycle or not), which we suppose could influence the muscular response, by the chemical

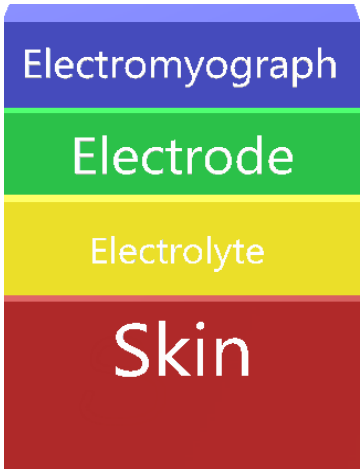


Fig. 3. Diagram of the Electromyograph connection. Own elaboration

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

elements that are in the process of communication as illustrated in figure 2. Their responses are shown in Table 1.

The experiment consists in performing weightlifting, of 3 sets of 12 repetitions with time intervals of 30 bits per second, where the first 10 seconds are left in basal mode (resting muscle), after which the first series of 12 repetitions, in the second 56 approximately the rest is performed, and the second series starts at approximately 01:12 minutes, rests and the third series starts at approximately 02:12 minutes. The samples taken were 4 for each woman, one sample for each week. The weight's dumbbells is seven pounds. The full time of the experiment is about 3:15 minutes.

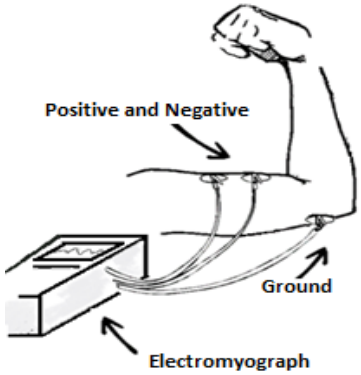


Fig. 4. Configuration to take electromyographic signals. Own elaboration

TABLE I
PEOPLE DESCRIPTION FOR THIS EXPERIMENT

Features	Woman 1	Woman 2	Woman 3	Woman 4	Woman 5	Woman 6	Woman 7
Age (year old)	22	33	26	21	23	24	25
Weight (kg)	60	68	63	67	94	89	69
Height (m)	1.73	1.62	1.60	1.66	1.68	1.73	1.67
Are you on period? (week 1)	Yes	No	No	No	No	No	No
Are you on period? (week 2)	No	No	Yes	No	No	No	Yes
Are you on period? (week 3)	No	Yes	No	No	No	Yes	No
Are you on period? (week 4)	No	No	No	Yes	Yes	No	No
Do you practice exercise?	Yes	Yes	Yes	No	No	No	No
Which sport do you practice?	Swimming	Spinning	Dumbbells and box	N/A	N/A	N/A	N/A
How many days do you exercise on a week?	2	5	5	N/A	N/A	N/A	N/A

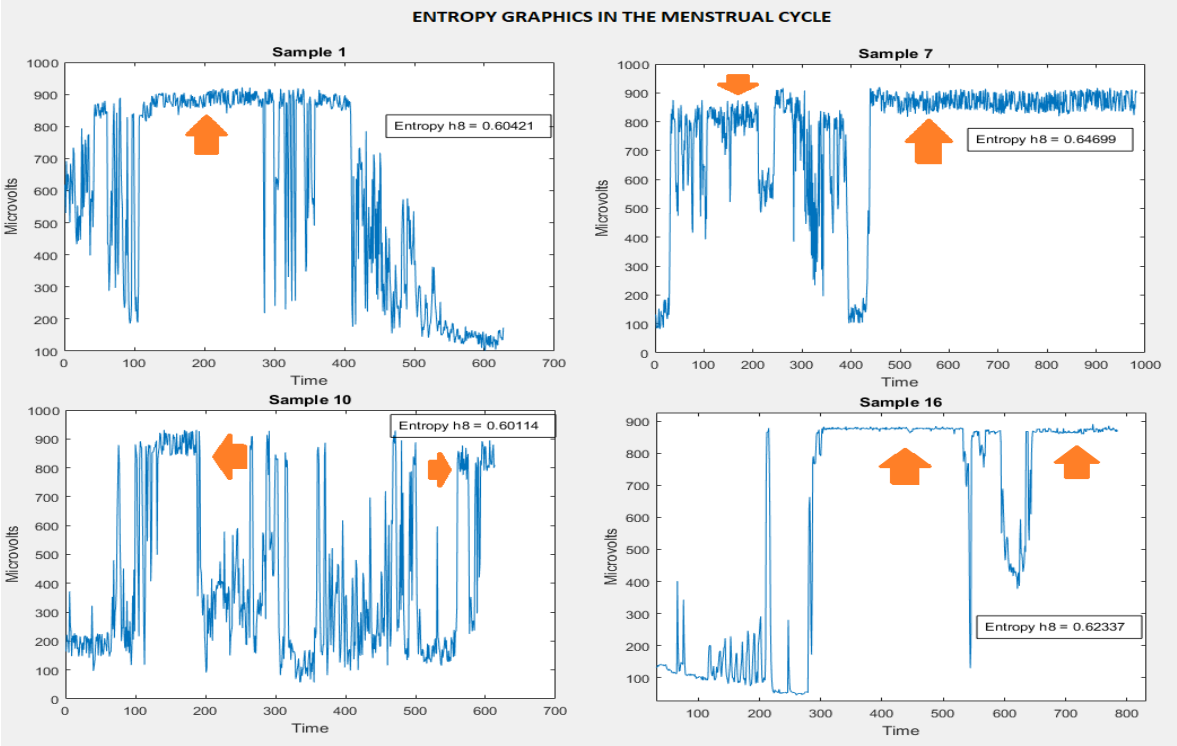


Fig. 5. Some graphics that shows the women period, and evaluation with entropy being 8 points of permutation.

Once the signal from each experiment is obtained, it was inferred that the physiological phase (woman's period) can be recognized in these signals (figure 5). This based on the information shown in table 1; the description of the chemicals that travel through the body and allow movement (membrane potential); the same signaling of it, the areas with orange arrows (figure 5) that mark the variations of the signal when the women were found in their period, unlike when they were not (figure

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

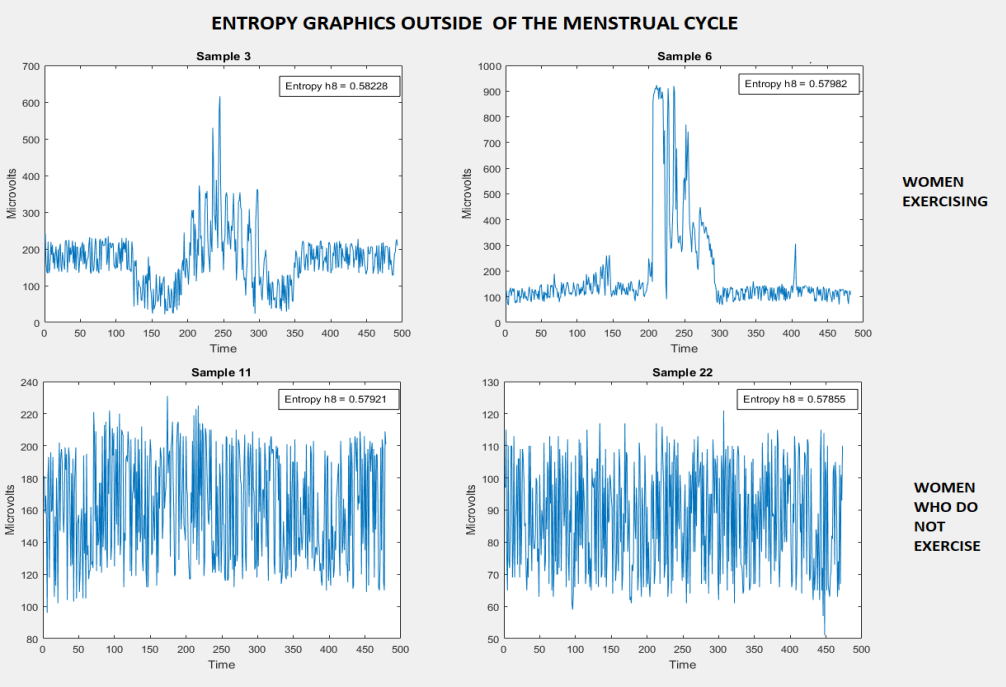


Fig. 6. Some graphics that shows the women without period, and evaluation with entropy being 8 permutation. In the first row are some women graphics exercising, meanwhile in the second row are some women graphics who do not exercise.

6); and to the application of the permuted entropy, in which 8 permutation points were used [13,15].

The entropy used allows the evaluation of nonlinear time series (chaotic series such as the case of the signals obtained in the muscle, figures 5-6), as well as allowing the comparison between signals, thus providing the necessary information to group the elements that belong to the same study.

The Hilbert Huang transform was also applied [20, 22-25], as another comparative method, where for this case study, where the woman is in her period, the forms of the signal are like rose petals, unlike when they are not in their period, the signals are replicated almost in the same way, as shown in figures 7-8. Finally, the Fourier transform [17-19] was applied, with which the enveloping characteristic of a biological signal mentioned by Hodgkin-Huxley in [24-25] is visualized.

IV. Conclusions and Result

It was found that in a series of biological time obtained from the muscular response can be modified by the phase of the

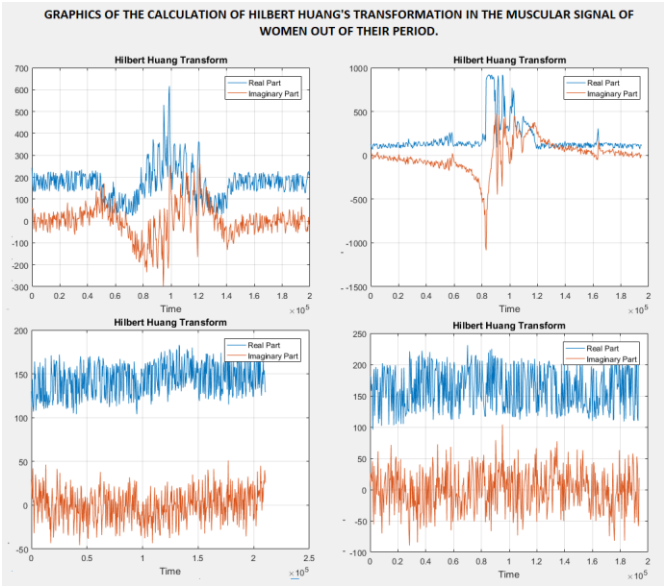


Fig. 7. Graphics of the calculation of Hilbert Huang's transformation in the muscular signal of women out of their period. The red signal is the result of the Hilbert Huang's Transform, and the blue signal is the bio-signal real.

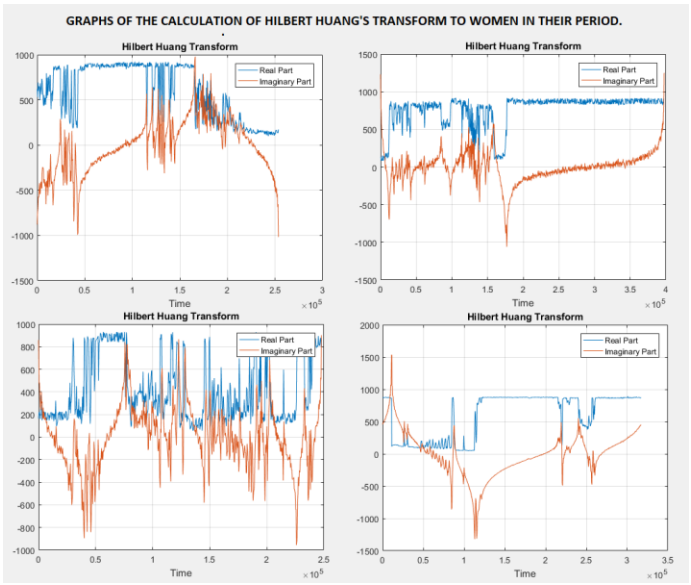


Fig. 8. Graphs of the calculation of Hilbert Huang's transform to women in their period. The red signal is the result of the Hilbert Huang's Transform, and the blue signal is the bio-signal real.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

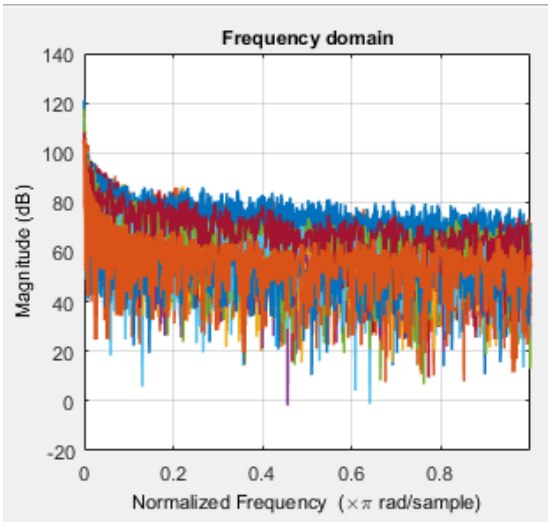


Fig. 9. All individual graphs of the Fourier transform of each bio-signal. Graph of the dominant frequency in the signs of the four weeks of the 7 women.

menstrual cycle, and that this is firstly clearly seen from the data obtained with the electromyography.

Later with the entropy we could know the level of uncertainty of the same signal, and that the singular behavior in the biological signal in the period of a woman, does not distinguish between whether or not exercise is done.

In the results we observed that when the woman is not in her period, between 0.5 - 0.587 entropy, and when she is in the menstrual cycle round between the 0.601 - 0.70 of entropy, always these evaluated with entropy of 8 permutations.

ACKNOWLEDGMENT

Authors would like to acknowledge to CONACYT, the Tecnológico Nacional de México, Instituto Tecnológico de León and our drawer Arlette Rodríguez; their support for this research.

REFERENCES

[1] GARCÍA, José María Artero; HURLÉ, J. Anatomía Y Fisiología Humana. Everest, (1978).
[2] SALAZAR, Osiel Arbeláez; ANGARITA, Jorge Gómez; VARGAS, Jairo Mendoza. Diseño De Un Electromiógrafo Con Procesador Digital De Señales Para Captación De Señales Musculares. Revista Médica De Risaralda, (2007), Vol. 13, No 1.
[3] H. A. Romo, J. C. Realpe And P. E. Jojoa, "Análisis De Señales EMG Superficiales Y Su Aplicación En Control De Prótesis De Mano," Vol. 14, No. 1, (2007).
[4] NEUMANN, Donald A. Cinesiología Del Sistema Musculoesquelético. Paidotribo, (2007).
[5] Jain, M. A Next-Generation Approach To The Characterization Of A Non-Model Plant Transcriptome. Curr. Sci. 101, 1435–1439 (2011).
[6] Ramakrishnan, S. & Karthick, P. A. Muscle Fatigue Analysis Using Surface EMG Signals And Time–Frequency Based Medium-To-Low Band Power Ratio. Electron. Lett. 52, 2–3 (2015).
[7] Al-Mulla, M. R., Sepulveda, F., Colley, M. & Kattan, A. Classification Of Localized Muscle Fatigue With Genetic Programming On Semg During Isometric Contraction. Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. Eng. Futur. Biomed. EMBC 2009 2633–2638 (2009). Doi:10.1109/IEMBS.2009.5335368
[8] Gerardo, J. Et Al. Polynomial Approximation Of Time Series Of Pupil Response To Controlled Light Stimuli. 7, 1–10 (2017).
[9] BRILLINGER, D. R. Time Series: Data Analysis And Theory, Society For Industrial Mathematics (2001).
[10] N. Bhadra, K. L. Kilgore, "High-Frequency Electrical Conduction Block Of Mammalian Peripheral Motor Nerve", Muscle Nerve, Vol. 32, No. 6, Pp. 782-90, Dec (2005).
[11] Liu, H. Steady State Membrane Potential And Sodium Current Changes During High Frequency Electrical Nerve Stimulation. 1006–1009 (2015). Doi:10.1109/ICMTMA.2015.245.
[12] Baxi, K. And Martin, R. Entropia. Architecture, Art, Design, Fashion, History, Photography, Theory And Things. Black Dog Pub. (2001).
[13] Bejan, A. Entropy Generation Minimization: The Method Of Thermodynamic Optimization Of Finite-Size Systems And Finite-Time Processes; CRC Press: Boca Raton, FL, USA, (1995).
[14] B. Christopher, P. Bernd. "Permutation Entropy – A Natural Complexity Measure For Time Series". Procedia Environmental Sciences. Vol. 88. (2002). Pp. 4102-4106.
[15] Qing, J., Bhatti, M. M., Abbas, M. A., Rashidi, M. M. & Ali, M. E. Entropy Generation On MHD Casson Nanofluid Flow Over A Porous Stretching / Shrinking Surface. Doi:10.3390/E18040123
[16] Mosquera-Dussan, O. L. & Botero-Rosas, D. A. Nonlinear Analysis Of The Electroencephalogram In Depth Of Anesthesia Análisis No Lineal De La Señal De Electroencefalograma En Profundidad Anestésica. 45–56 (2015). Doi:10.17533/Udea.Redin.N75a06
[17] E.O. Brigham, The Fast Fourier Transformi, Prentice-Hall, Englewood Cliffs, N. J., (1974).
[18] A. Oppenheim Y R. Schafer, Discrete-Time Signal Processing, Cap. 8, 9 Y 12, Prentice-Hall Inc., (1989).
[19] W.T. Cochran, What is the Fast Fourier Transform?, IEEE, Trans. Audio Electroacoust., Vol. AV-15, (1967), pp. 45-55
[20] 1. Ding, H., Huang, Z., Song, Z. & Yan, Y. Hilbert – Huang transform based signal analysis for the characterization of gas – liquid two-phase flow. 18, 37–46 (2007).
[21] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of Royal Society London Series A (1998);454:903–95.
[22] Balocchi R, Menicucci D, Varanini M. Empirical mode decomposition to approach the problem of detecting sources from a reduced number of mixtures, EMBS. In: Proceedings of the 25th annual international conference of the IEEE, vol. 3. (2003). p. 2443–6.
[23] Wu ZH, Huang NE. A study of the characteristics of white noise using the empirical mode decomposition method. Proceedings of Royal Society London Series A (2004); 460:1597–611.
[24] Jane Cronin, Mathematical Aspects of Hodgkin-Huxley Neural Theory, Cambridge Studies in Mathematical Biology, Cambridge University Press, 1987; ISBN: 0521334829,9780521334822.
[25] Hodgkin A.L., Huxley A.F. A Cuantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. J Physil. 117, 500-544 (1952)

Prostate Near-Infrared Fluorescent Image Segmentation Using Hopfield Neural Network Classifier

Rachid Said Sammouda
Computer Science Department
College of Computer and Information Sciences, King Saud University
Riyadh, KSA
dr.rachidsammouda@gmail.com, rsammouda@ksu.edu.sa

***Abstract*— Thousands of people die every year due to prostate cancer. The prognosis of prostate cancer is very slow in most of the cases, but it can cause the death of the patient. The diagnostic pattern and the strategies of health care systems have changed over the last ten years. This change occurred rapidly due to the easy availability and an outburst in patient's data. This data is used as input data to Computer-Aided diagnosis systems. The objective of this research is to improve the diagnosis by developing a prototype arrangement for revealing, detecting and classifying the prostrate tumor. This is achieved by using Near-infrared and Mid-infrared spectrums of prostate pathological images. This optical imaging technique is a potent tool for cancer investigation that relies on stimulating endogenous chromophores or applying contrast agents able to target cancer cells. Here, we present a segmentation method of images obtained using PSMA (Prostate Specific Membrane Antigen) targeting optical imaging probes for NIRF (Near Infrared Fluorescence). This phenomenon is applied for intra-operative visualization of prostate cancer. An Artificial Neural Network classifies the pixels into distinguished clusters. Preliminary tests were conducted. The outcomes of these tests reveal that the planned segmentation technique can enhance the existing clinical practice in identifying prostate area. According to the NIRF image, shape and volume analysis could be conducted using the segmentation result for further investigations.**

Keywords- Hopfield Neural Network Classifier; Near-Infrared Fluorescence optical images; Prostate Cancer; PSMA (Prostate Specific Membrane Antigen); Segmentation;

I. INTRODUCTION

Cancer or malignant tumor occurs due to the abnormal growth of the cells. The prognosis of cancer occurs due to the movement of cancerous cells in the body by using mediums i.e. the blood and lymph. These cancer-cells attack the healthy cells and destroy them. The cancer cell grows by cell division process causes angiogenesis, i.e. formation of new blood vessels. For global public health, cancer has become a major risk factor. Regardless of the progress in wide-ranging therapy, cancer is a straining financial difficulty for patients in all societies. The detection of cancer is very important at its earliest stages. This early detection is very difficult because of the reduced level of specificity and sensitivity regarding current diagnostic approaches of imaging. There are different types of cancer; prostate cancer is one of them [1, 2].

Prostate cancer is the cancer that occurs in the tissues of the prostate gland. The function of prostate gland is the production of seminal fluid. The seminal fluid is required for the nourishment and transportation of sperms. Prostate cancer lives as it is born slow-growing and benign or fast-growing and dangerous 3. The early stage diagnosis of cancer is very important to prevent its prognosis [3]. For this purpose, the research is going on to device new techniques for early diagnosis and detection.

Clinical organizations are working on the prevention and treatment of the cancer. Furthermore, different strategies are planned to improve the diagnostic methods. The new aim is to develop the non-invasive methods such as imaging method. The biomedical imaging devices are used very frequently today, while further developments are underway to produce more advanced apparatus. These devices work at the cellular, molecular or tissue levels and make the diagnosis more accurate and favorable. The studies at the molecular and cellular levels help to know the mechanism of the prognosis of cancer. In addition to this, the patients prefer the non-invasive method of diagnosis. Keeping all these points in mind, imaging methods are becoming common and advancements in imaging modalities are progressing. The appropriate usage of near infrared segmentation is also a new approach for diagnosing the condition of prostate cancer. The studies are carried out on mice to evaluate the affectivity of the proposed method [4, 5].

The extensively used method for the detection of the pathological changes occurring due to cancer is imaging modalities. The examples of these widely used imaging methods are CT, PET, ultrasound, and MRI. Such methods show results in cases of benign lesions. However, in malignancies the imaging technique fails to get a clear contrast between the benign and the malignant. Moreover, the adjacent normal tissues add further confusion. To improve the detection and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

examination of cancer at any phase, it is very essential to develop a high contrast narrative imaging method to augment the diagnosis and therapeutics [1].

Prostate cancer and breast cancer are two frequent types of cancer and their diagnosis has become a challenge for the scientists. The difficulty in localization of cancer cells is the main barrier. Furthermore, it is very difficult to differentiate between the normal and tumor cells. New strategies are designed to localize the cancer cells in imaging. These strategies comprises of the techniques of labeling methods. The ligands which are tumor-specific and having sympathetic pharmacokinetics are developed for labeling purpose [1].

The method of capillary permeability and in vivo tumor growth selectively increase the expression of tumor markers, and tumor delivery changes. Favorable pharmacokinetic, and small tumor markers developed using pre-targeting strategies are important in improving the diagnostic approach [2]. In diagnosing cancer, optical imaging by Near-infrared Fluorescence (NIRF) is a dominant trend. It relies on activating endogenous chromophores or applying contrast agents that can target cells. Several new NIRF agents have been developed including heptamethine carbocyanine dyes. Some of these agents have become commercially available in recent years, such as Cy5.5 [4] and IR Dye 800-CW [5]. These have been coupled with peptides or antibodies and successfully used for the targeted visualization of neoplastic tumors in animal models [6]. Xinning et al. [7] and others [8-12] have developed optical molecular imaging approaches to differentiate between tumors and surrounding normal tissues during surgery, for reviews see [13-15]. A first-in-human-study has been conducted for ovarian cancer [16], indicating progress in this field. The clinical and medical functions of these new NIRF agents offer great promise for future.

The over expressed antigen in most of prostate cancer is PSMA [17-22]. This is the reason that it is a useful biomarker for discrimination of prostate cancer tissue from surrounding normal tissues. Prostate tumor expressing the PSMA receptors were implanted into the flank of mice as previously described [23]. Control tumors that did not express the PSMA receptor were implanted on the opposite flank. When tumors reached the appropriate size the mice were administered a ligand for PSMA labeled with a fluorophore for detection by fluorescence imaging.

II. Generation of Data

A. Mouse Tumor Xenograft Models

Animals were observed every other day until tumors reached about 10 mm in diameter. Orthotropic implantation of prostate cancer was carried as previously described. Briefly, six to eight weeks old male nude mice lacking thymus gland were anesthetized. The composition of anaesthetic solution was 5 mg/mL ketamine/ 3 mg/mL xylazine solution in 0.9% saline and the volume given was 200 uL. The route of administration was intra-peritoneal (in the peritoneal cavity). The lower abdomen was open to expose the dorsal-lateral prostate, to which 10 to 20 uL cell suspension in PBS (5x107 cells/ml) was injected. The incision in the abdominal wall was closed. After four weeks, animals were ready for experimentation.

B. In-vivo NIR Imaging Studies

With the assistance of Maestro in- vivo Imaging system (Perkin-Elmer, Waltham, MA), imaging was performed. 1 nmol of NIR probe in PBS through tale vein injection was given to each mouse. Imaging was carried out by using the appropriate filter set (deep red filter set for PSMA-1-IR800 and yellow filter set for PSMA-1-Cy5.5). Different points were selected for imaging. The temperature of 37oC was tuned for the imaging bed during imaging. A nose cone was adjusted with imaging bed for inhalation of isofluorane. Cervical dislocation was used to sacrifice mice after imaging mice over 5 days post injection. To perform ex vivo imaging, harvesting of tissues, for example kidneys, heart bladder and liver was done.

Fluorescent molecular tomographic (FMT) images were obtained using the FMT2500 device (Perkin-Elmer, Waltham, MA) and three-dimensional reconstructions of fluorescent signals were acquired using the accompanying software, TrueQuant. Quantification of fluorescent signals was obtained by calibration of PSMA-1-IR800 and PSMA-1-Cy5.5 using the 780 nm and 680 nm channel respectively. To block the binding of PSMA-1-NIR in mice, mice were co-injected with 1 nmol of PSMA-1-NIR probes and 100 nmol of ZJ-MCC-Ahx-YYYG, an analogue of PSMA-1 with similar binding affinity but with no optical probe attached.

Maestro Imaging System and FMT were the two imaging methods that were used to image mice for up to 24 hours. For orthotopic mouse models, mice were imaged at 4 hours or 24 hours by using Maestro Imaging System. 1 nmol of PSMA-IR800 injection at 4 hours or 1 nmol of PSMA-1-Cy5.5 at 24 hours were injected in post tail vein. After the completion of the optical imaging, the mouse was euthanized, the abdomen was opened to expose the tumor, and the mouse was again imaged. Finally, tumor was harvested for ex vivo imaging.

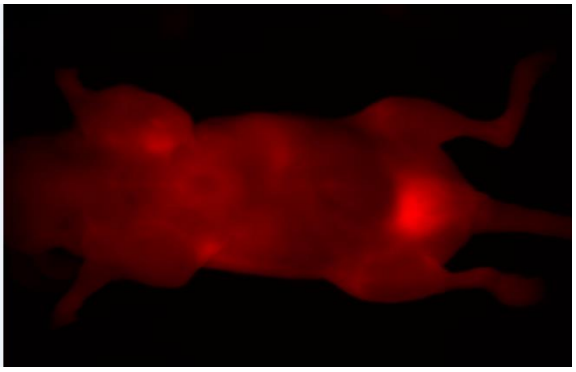


Figure 1. Shows a NIRF sample image of mice model with prostate cancer.

Figure1 shows a NIRF image of prostate tumors obtained by using targeted imaging probe of Prostate specific membrane antigen in a mouse model. Several similar NIRF images were collected in our previous study which has been conducted to develop PSMA-targeted near infrared (NIR) optical imaging probes. These were used for visualization of prostate cancer intra-operatively. A high affinity PSMA ligand (PSMA-1) was

National Plan for Science and Technology at King Saud University (KSU)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

synthesized with low molecular weight, and further labeled with commercially available NIR dyes: IRDy800 and Cy5.5 [5]. It demonstrated the utility of such probes to selectively bind to prostate tumor in vivo targeting both heterotopic and orthotropic prostate tumors. A challenge for these types of studies is to correctly interpret the imaging data to accurately reflect the margin of the cancer. In cancer research, it is very difficult to obtain reproducible, accurate, precise and intent assessment. The dilemmas occur due to the variability of personnel, biological dissimilarity, and natural unpredictability. NIRF imaging technique identifies the cancer by providing fluorescent information from every pixel in the image. For prostate cancer, the NIRF imaging CAD system (Computer Aided Diagnosis) could be classified and analyzed to build a set of sharp diagnostic rules. We present a segmentation method of the NIRF images as the first and bottleneck entity of the CAD system in the next section

III. Segmentation Method

All of the above described phenomena were studied to determine the importance and use of infrared radiation for obtaining a better health approach [24]. To screen medical imaging, segmentation of image is imperative. A progressive method used for screening purposes in the last few decades is fuzzy segmentation method. The broadly used fuzzy method is based on c-means algorithm. The accomplishment of introducing fuzziness for each image pixel is successful as it is fit for images. The fuzziness promotes the bunching and clustering of the image pixels. This method assists in preserving more information in cluster form. The original image obtained has hard and crisp segmentation process which does not give precise information. That is why the clustering method is preferred [25].

Similar to that, we have used the Unsupervised Hopfield Neural Network Classifier (UHNNC) in segmentation of different types of medical and natural color images [26, 27]. The segmentation results have been appreciated with respect to the multi-dimensionality of the data type used for segmentation. This means the UHNNC gives better segmentation results as far as getting more information about the pixel of the scene under segmentation.

A grid of NM neurons is present in UHNNC architecture. The rows and columns are well defined. In NM grid, the alphabet N is used to show the size of the image; whereas the numbers of the cluster formed are represented by M. Columns are used to characterize a class while pixels are represented in row form. The network is deliberate to sort the area of the features themselves.

By using a distance scale, compactness of each category is calculated. The problem of Segmentation is considered as a partition of N pixels of P characteristics among M clusters or so that the cost of energy (errors) function can be minimized by the tasks of pixels:

$$E = \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{M} R_{kl}^{n} V_{kl}^{2} \quad (1)$$

The similarity distance is measured and represented by R_{kl}. It shows the distance between k[th] pixel and the centroid of class *l*. It is given as:

$$R_{kl} = \|X_k - \bar{X}_l\| \quad (2)$$

In the above equation X_k represents the P-features vector, for color images, P=3 in the RGB color space, as k[th] pixel's intensities, while $\bar{X}_l$ is the class l's centroid, and is shown as:

$$\bar{X}_l = \frac{\sum_{k=1}^{N} X_k V_{kl}}{n_l} \quad (3)$$

To allocate a label m to the pixel, the input-output function for the k^{th} row, winner-takes-all learning is used by HNN. It is given by:

$$\begin{cases} V_{kl}(t+1) = 1, \ if \ \ U_{kl} = Max\{U_{kl}(t), \forall\, l\} \\ V_{kl}(t+1) = 0; \qquad\qquad otherwise \end{cases} \quad (4)$$

UHNNC is used for the minimization purpose and by working out a group of equations of motion, the resultant obtained is:

$$\frac{\partial U_i}{\partial t} = -\mu(t)\frac{\partial E}{\partial V_i} \quad (5)$$

U_i , represents the input of i^{th} neuron, while the output is represented by symbol V_i. For increasing the convergence speed of the HNN, $\mu(t)$ is used as a scalar positive function of time that we have defined and verified in our study its efficacy in assuring and forcing the network to converge after a pre-specified time T_s as follows:

$$\mu(t) = t * (T_s - t) \quad (6)$$

A group of neural dynamics is obtained by relating the equation (5) to equation (1) and is given by:

$$\frac{dU_{kl}}{dt} = -\mu(t)(R_{kl}^{n} V_{kl}) \quad (7)$$

The UHNNC segmentation algorithm can be easily summarized as:

Phase1. The neurons' input is initialized to randomly assigned values.

Phase2. The new output value for every neuron can be obtained by applying the input-output relation given in (4).

Phase3. With respect to the equation (3), centroid can be computed for each class.

Phase4. The input of each neuron is required to be updated by solving the set of differential equation in (5) and given as:

$$U_{kl}(t+1) = U_{kl}(t) + \frac{dU_{kl}}{dt} \quad (8)$$

Phase5. Loop to Phase2 with T_s times.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. SEGMENTATION RESULTS

The main objective regarding the preliminary research is to reveal the importance of the anticipated approach of segmentation for the diagnosis of prostate cancer by using NIRF imaging technique. However, although optical imaging by NIRF in cancer researches is an influential research instrument as mentioned above, it remains a one dimensional information set about the scene's environment and its images' segmentation using UHNNC which is of limited contrast.

To overcome that contrast limitation, we have produced an artificial multidimensionality of the NIRF image using dependent chromatic redundancy in the RGB color space. Figure 2 (a) shows the NIRF sample image of mice model with prostate cancer of Figure1.

The Green and Blue channels obtained by redundancy from Figure 2 (a) are shown in (b) and (c) parts of Figure 2. The (d) part of Figure 2 regarding the RGB color space shows the full color display of the three components. The above described UHNNC is applied to several NIRF images. The results show that segmentation of most of the images can be obtained successfully by using our algorithm. The segmentation has clearly distinguishable areas as background, or other uniform clusters with respect to their features in the input images.

Figure 3 shows the segmentation result using the UHNNC of the NIRF image of a mice model Figure2 (a) and its two redundant green and blue color filters, Figure 2 (b) and (c), with respect to number of clusters, 3, 4, 5, and 6, respectively to (a), (b), (c), and (d).

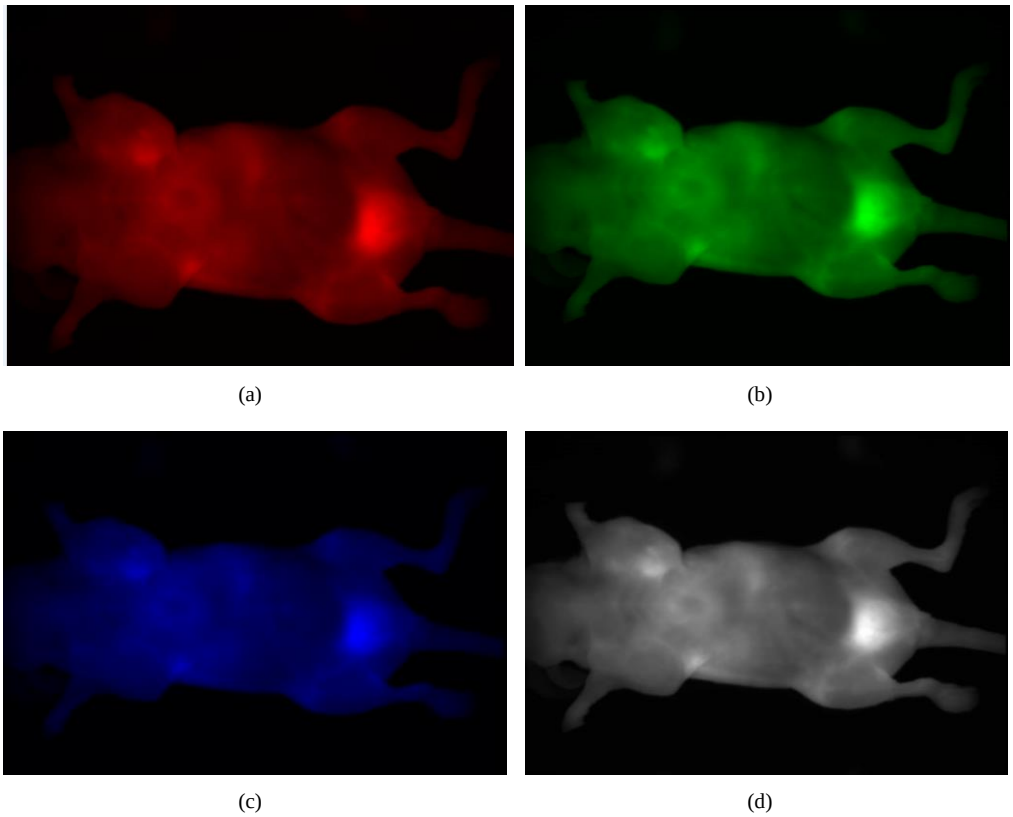


Figure 2. (a) is the raw NIRF image of the case under study, the same image in Figure 1, (b) and (c) are green and blue component obtained by redundancy from (a), and (d) is the full color display of (a), (b), and (c) with respect to the RGB color space.

The curves of the segmentation problem energy function are shown in Figure 4. The segmentation problem during its optimization using the described UHNNC with respect to the number of cluster L is decided by the user based on anatomical, medical information.

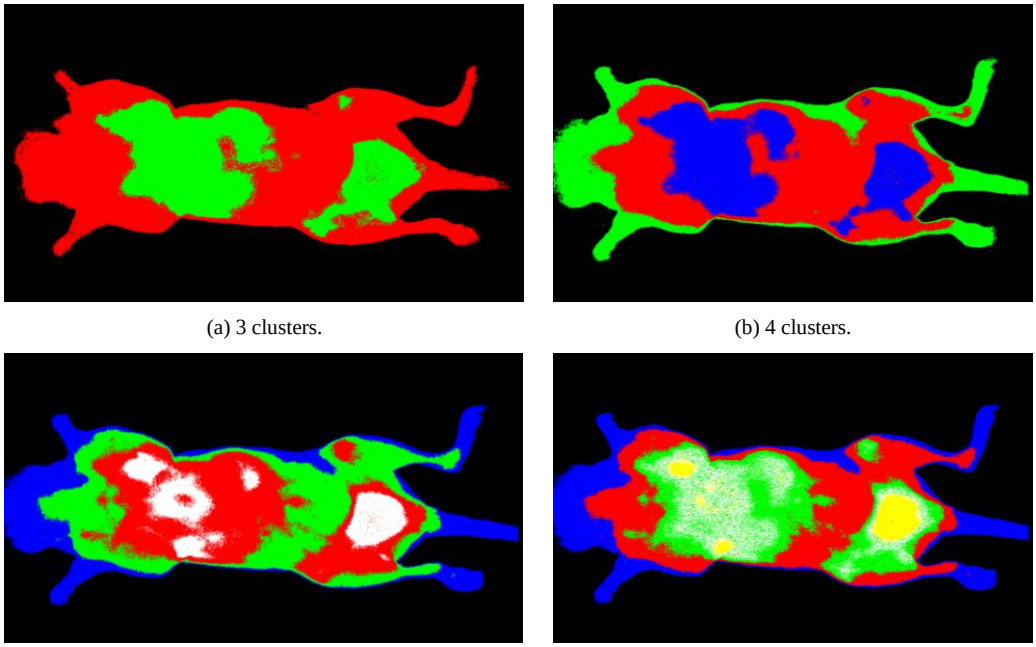
We realize that the prostate region starts to appear as independent region with its outer borders when clusters number is equal to five.

Figure 6 (d) obtained with six clusters, shows the prostate region with an outer and inner regions.

Figure 5 shows more specific regions within the prostate area, for which we have conducted segmentation with more clusters, eight, nine and ten. The convergence optima of the UHNNC with respect to the number of clusters used during the segmentation process is shown in Figure (6).

As can be seen, the UHNNC has reached better local optima when used with ten clusters, as there are more intensity variations among pixels, however, the prostate region, cluster number 4 in Figure 7, remains among the cluster of the highest mean value.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(a) 3 clusters. (b) 4 clusters.

(c) 5 clusters. (d) 6 clusters.

Figure 3. shows the segmentation result using the UHNNC of the NIRF image of a mice model Figure2 (a) and its two redundant green and blue color filters, Figure 2 (b) and (c), with respect to number of clusters, 3, 4, 5, and 6, respectively to (a), (b), (c), and (d).

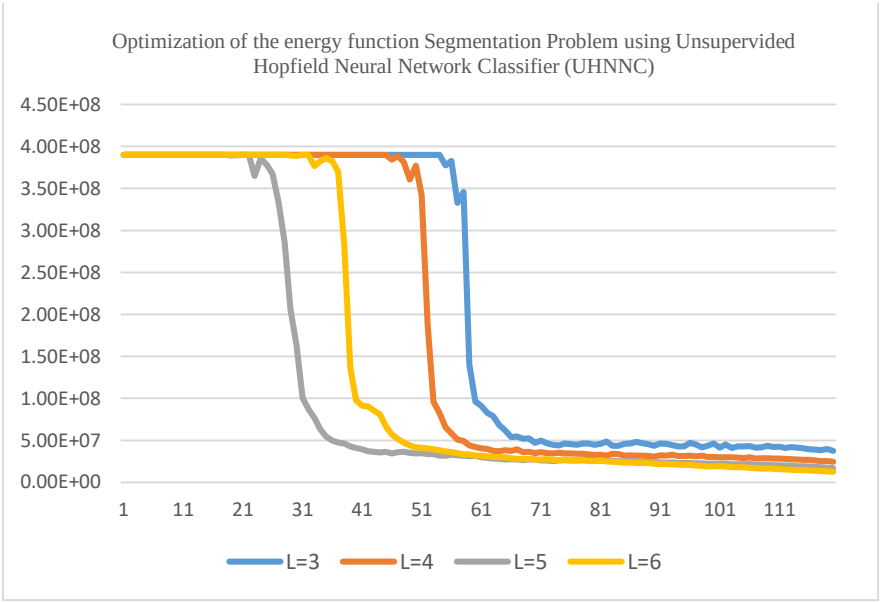
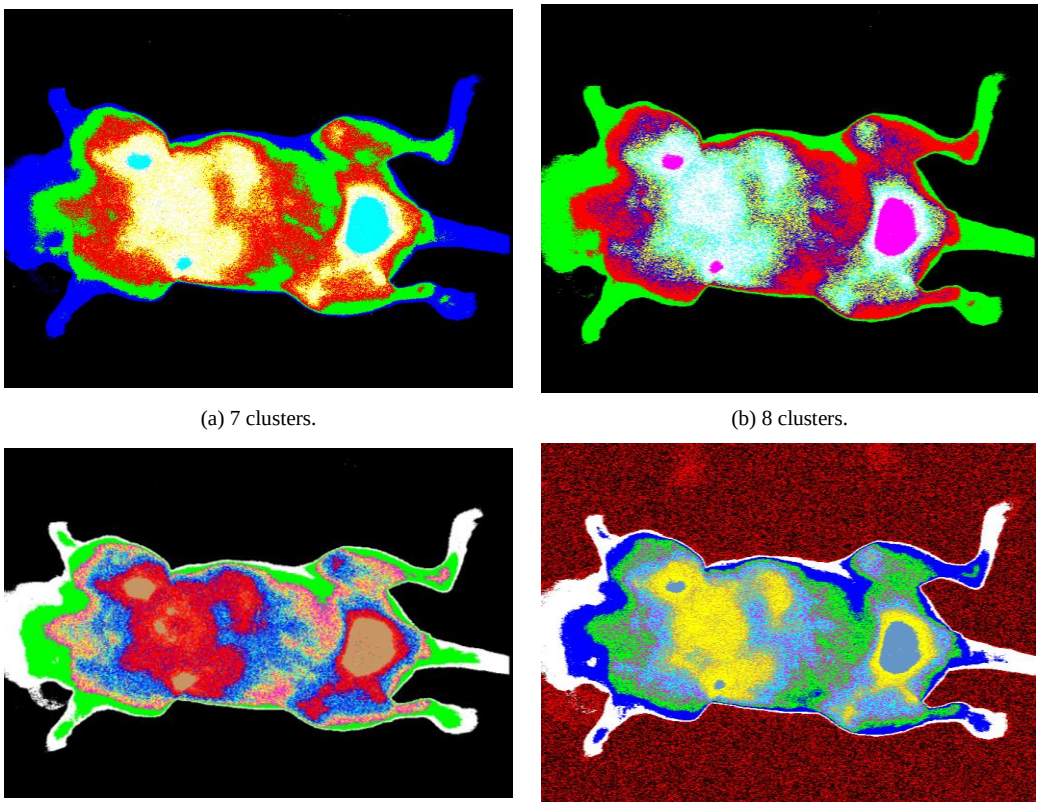


Figure 4. shows the curves of the segmentation problem energy function during its optimization using the here described UHNNC with respect to the number of cluster, L, decided by the user.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(a) 7 clusters. (b) 8 clusters.

(c) 9 clusters. (d) 10 clusters.

Figure 5. shows the segmentation result using the UHNNC of the NIRF image of a mice model Fig. 2 (a) and its two redundant green and blue color filters, Fig. 2 (b) and (c), with respect to number of clusters, 7, 8., 9 and 10, respectively to (a), (b), (c), and (d).

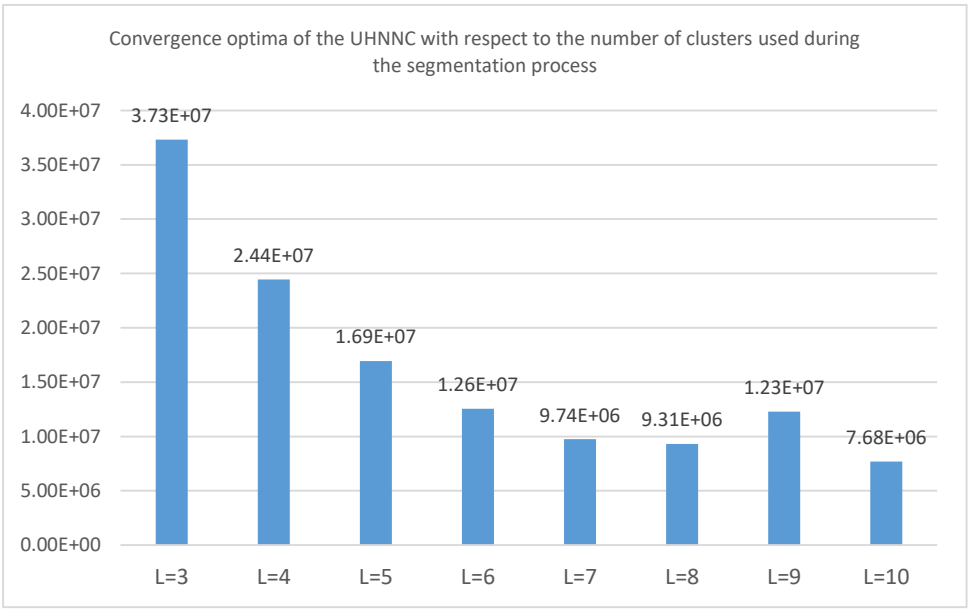


Figure 6. shows the convergence value of the energy function of the UHNNC during the segmentation process, of the NIRF image shown in Fig. 1, with respect to the number of clusters decided by the user.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

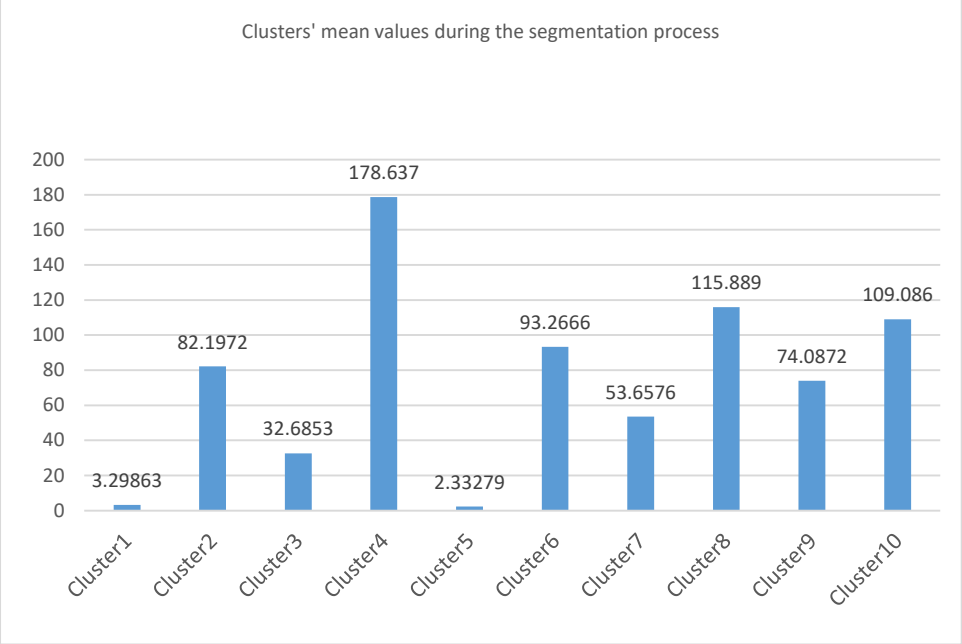
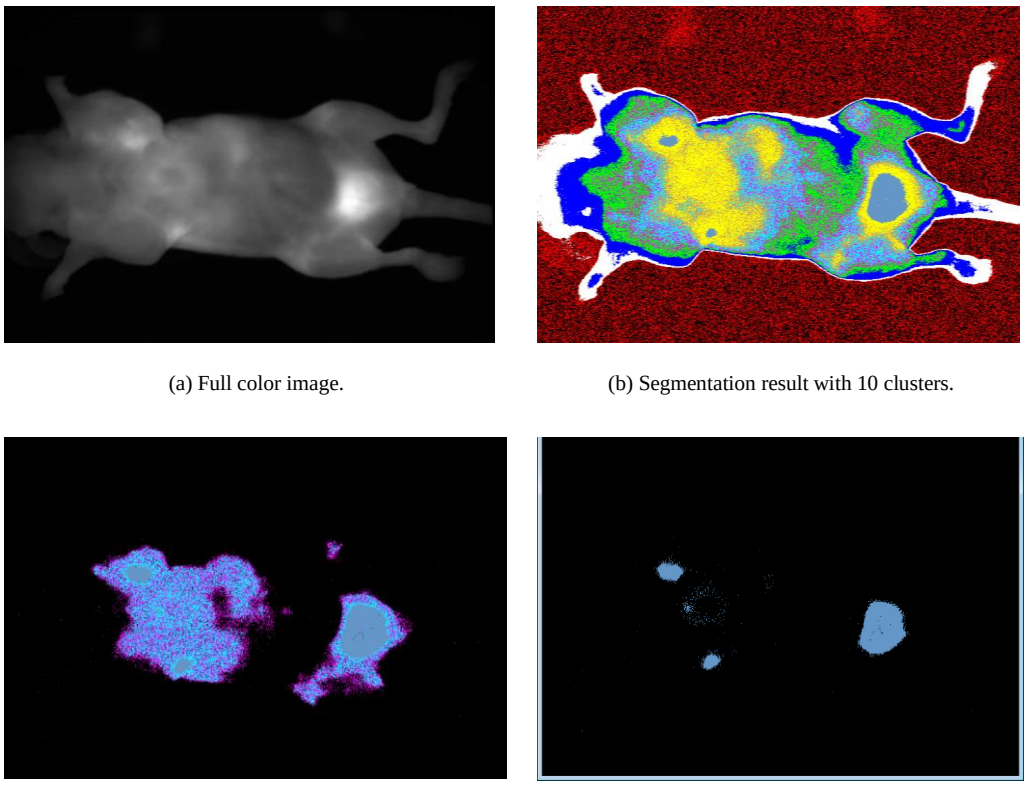


Figure 7. shows the mean value of each cluster in the segmentation result of the NIRF image shown in Fig. 1, with 10 as number of clusters, decided by the user.



(a) Full color image.

(b) Segmentation result with 10 clusters.

(c) Clusters with the first three maximums of mean values.

(d) The cluster having the maximum mean value.

Figure 8. shows the full color image of the case under study in (a), its corresponding segmentation result with 10 clusters in (b), the clusters of the first three maximum mean values in (c) and the cluster with the maximum mean value, including the prostate region, in (d).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 8 shows the details of the full color image of the case under study in (a). While in (b) the result with 10 clusters of corresponding segmentation is shown. The clusters of the first three maximum mean values in (c) and with the maximum mean value, including the prostate region, is shown in (d).

We can realize here that even with more clusters, the prostate region remains as an entity region with lower intensity variation, and did not split into two clusters as the background. The latter will be used as a mask to take out the section of attention from the unrefined image for further diagnosis and design of the CAD system for prostate cancer diagnosis. All these figures make the procedure of segmentation of near-infrared fluorescent easy to understand.

V. Discussion

The imaging obtained by using near infrared is very helpful for the diagnosis as well as for the surgical approach. It provides accurate images of the cancer cells that differ from the images of normal cells. The differentiation makes the diagnosis easy and painless. The differentiation marked by using this technique aids the dissection and categorization of the tumor related cells without developing any harmful effect [16].

This near-infrared imaging technique not only acts as a diagnostic tool, but also traces the response of the cells to the chemotherapeutic agents. For the therapy of prostate cancer, it is very important to formulate a drug which has high therapeutic efficacy and fewer side effects. This is done by taking the measurements of the images. The measurements are taken to find out the reduction in the size of tumor. This, however, is a lengthy process. Despite long time delay, this technique is considered as an important indicator for the trial of new drugs. The efficacy of a new drug molecule can be determined by using this approach. Thus this tool is helpful in selecting the most significant and effective treatment [27].

One of the recent modality of imaging is hyper-spectral imaging. It is a spectroscopic method, and the data obtained from this imaging method is utilized for non-invasive approach in cancer detection. The differentiation of tumor cells from the healthy cells is necessary. This is done by quantitative analysis. For prostate cancer, the data is obtained by use of an advanced image data. The analysis of hyper-spectral image is done to obtain the data which is utilized for the detection of cancer. For the purpose of differentiation of normal and cancerous cell, the spectrum was taken out for both kinds of cells. The studies were conducted in order to detect prostate cancer on the mice having tumor. Moreover, pathological slides were also used for detection. By using this technique, the images of normal and tumor cells were taken and the reflectance properties of both cells were extracted. These images showed that the reflectance properties of both cells are different. The sensitivity and specificity of this method are fine. By using the data obtained by spectral images, is very helpful to differentiate between normal and tumor cell, so the safe dissection of malignant areas is possible [28]. The determination of the in vivo cell death is possible by the use of near infrared fluorescent method. For this purpose, fluorescent probes are used. For example, active Cy-annexin is used in non-radioactive techniques. NIRF probe having active cy-annexin is used to determine the anti-proliferative properties of the molecules which are used as chemotherapeutic regimens. By analyzing the properties of the regimens, it is very easy for the clinicians to choose the chemotherapeutic agents for the prostate cancer [28].

Another study indicates the importance of the near-infrared spectroscopy for the diagnosis and detection purpose. The near infrared image segmentation has achieved the significance because of its non-invasive property. Due to this property, the technology is widely used. The diagnostic markers are used to detect the chromophores difference [29].

The endogenous chromophores of normal cells are different from the cancer cells, and this quality is used to identify the cancer cells. This detection is based on near-infrared radiations and the biomarkers (for example, lipids bands, deoxy-haemoglobin, oxy-haemoglobin and water bands etc). In addition to NIR, different agents are used to increase the contrast of the image [30]. All these studies indicate that near infrared rays fluorescent imaging is a very important method for the detection of the cancer cells. The segmentation of the image obtained by this technique is an advanced approach for this procedure. This new innovation is very promising and it can be used as a potential aid in the war against prostate cancer. In the future work, we will apply the method proposed in [31] for NIRF images de-noising before segmentation them by the previous proposed methods in [32-34] and compare the results with the proposed method in this paper.

VI. Conclusion

In this paper, we presented the use of NIRF images for prostate cancer diagnosis. NIRF technology has been considered widely for biomedical research and clinical application since it has been demonstrated that near-infrared is an appropriate optical opening for profound tissue imaging.

As shown in the sample of NIRF images, used in this study, these dyes show the ability of providing fine information and behaviors about different mice's tissues. The finest information is utilized to develop the new strategies of diagnosis and detection. The segmentation of these NIRF images using our modified UHNNC confirms the fact that prostate cancerous tissue takes more fluorescent material than normal tissue.

The analysis process conducted among the different clusters, of the segmentation results; prove the low intensity variation among pixels of the cancerous tissue. This makes the prostate cancerous region presented by smooth region and sharp edges. In our future work, we will use these features in order to extract automatically the region of interest (ROI) as prostate tissue, and focus on the internal behavior of its cells for better guidance in prostate cancer therapy and early diagnosis.

Acknowledgment

The author is appreciative to National Plan for Science and Technology for their assistance in this research at King Saud University (KSU), Riyadh, Kingdom of Saudi Arabia (KSA), under project number: 10-Bio-1905, and the hosting of the assignment provided by the College of Computer and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Information Sciences at KSU. Also, the author is appreciative to James P. Basilion researcher at the Department of Radiology and NFCR Center for Molecular Imaging, Case Western Reserve University, and Cleveland, OH, USA for his support as consultant in this project.

REFERENCES

[1] X. Yi, F. Wang, W. Qin, X. Yang and J. Yuan, "Near-infrared fluorescent probes in cancer imaging and therapy: an emerging field", International Journal of Nanomedicine, p. 1347, 2014.

[2] R. Weissleder, C. Tung, U. Mahmood and A. Bogdanov, "In vivo imaging of tumors with protease-activated near-infrared fluorescent probes", *Nature Biotechnology*, vol. 17, no. 4, pp. 375-378, 1999.

[3] D. Pendick, "Prostate cancer lives as it is born: slow-growing and benign or fast-growing and dangerous - Harvard Health Blog", *Harvard Health Blog*, 2017. [Online]. Available: http://www.health.harvard.edu/blog/prostate-cancer-lives-as-it-is-born-slow-growing-and-benign-or-fast-growing-and-dangerous-201308146604. [Accessed: 21- Dec- 2017].

[4] S. Banerjee, M. Pullambhatla, Y. Byun, S. Nimmagadda, G. Green, J. Fox, A. Horti, R. Mease and M. Pomper, "68Ga-Labeled Inhibitors of Prostate-Specific Membrane Antigen (PSMA) for Imaging Prostate Cancer", *Journal of Medicinal Chemistry*, vol. 53, no. 14, pp. 5333-5341, 2010.

[5] Y. Chen, S. Dhara, S. Banerjee, Y. Byun, M. Pullambhatla, R. Mease and M. Pomper, "A low molecular weight PSMA-based fluorescent imaging agent for cancer", *Biochemical and Biophysical Research Communications*, vol. 390, no. 3, pp. 624-629, 2009.

[6] J. Wu, D. Pan and L. Chung, "Near-infrared fluorescence and nuclear imaging and targeting of prostate cancer", *Amepc.org*, 2017. [Online]. Available: http://www.amepc.org/tau/article/view/2764/3637. [Accessed: 21- Dec- 2017].

[7] J. Cutter, N. Cohen, J. Wang, A. Sloan, A. Cohen, A. Panneerselvam, M. Schluchter, G. Blum, M. Bogyo and J. Basilion, "Topical Application of Activity-based Probes for Visualization of Brain Tumor Tissue", *PLoS ONE*, vol. 7, no. 3, p. e33060, 2012.

[8] G. Blum, S. Mullins, K. Keren, M. Fonovič, C. Jedeszko, M. Rice, B. Sloane and M. Bogyo, "Dynamic imaging of protease activity with fluorescently quenched activity-based probes", *Nature Chemical Biology*, vol. 1, no. 4, pp. 203-209, 2005.

[9] G. Blum, G. von Degenfeld, M. Merchant, H. Blau and M. Bogyo, "Noninvasive optical imaging of cysteine protease activity using fluorescently quenched activity-based probes", *Nature Chemical Biology*, vol. 3, no. 10, pp. 668-677, 2007.

[10] Q. Nguyen, E. Olson, T. Aguilera, T. Jiang, M. Scadeng, L. Ellies and R. Tsien, "Surgery with molecular fluorescence imaging using activatable cell-penetrating peptides decreases residual cancer and improves survival", *Proceedings of the National Academy of Sciences*, vol. 107, no. 9, pp. 4317-4322, 2010.

[11] Y. Urano, M. Sakabe, N. Kosaka, M. Ogawa, M. Mitsunaga, D. Asanuma, M. Kamiya, M. Young, T. Nagano, P. Choyke and H. Kobayashi, "Rapid Cancer Detection by Topically Spraying a -Glutamyltranspeptidase-Activated Fluorescent Probe", *Science Translational Medicine*, vol. 3, no. 110, pp. 110ra119-110ra119, 2011.

[12] M. Veiseh, P. Gabikian, S. Bahrami, O. Veiseh, M. Zhang, R. Hackman, A. Ravanpay, M. Stroud, Y. Kusuma, S. Hansen, D. Kwok, N. Munoz, R. Sze, W. Grady, N. Greenberg, R. Ellenbogen and J. Olson, "Tumor Paint: A Chlorotoxin:Cy5.5 Bioconjugate for Intraoperative Visualization of Cancer Foci", *Cancer Research*, vol. 67, no. 14, pp. 6882-6888, 2007.

[13] S. Keereweer, J. Kerrebijn, P. van Driel, B. Xie, E. Kaijzel, T. Snoeks, I. Que, M. Hutteman, J. van der Vorst, J. Mieog, A. Vahrmeijer, C. van de Velde, R. Baatenburg de Jong and C. Löwik, "Optical Image-guided Surgery—Where Do We Stand?", 2017. [Online]. Available: http://link.springer.com/article/10.1007/s11307-010-0373-2#page-1. [Accessed: 21- Dec- 2017].

[14] U. Mahmood and R. Weissleder, "Near-Infrared Optical Imaging of Proteases in Cancer", *Molecular Cancer Therapeutics*, 2017. [Online]. Available: http://mct.aacrjournals.org/content/2/5/489.short. [Accessed: 21- Dec- 2017].

[15] "Molecular Imaging | Radiology", *Pubs.rsna.org*, 2017. [Online]. Available: http://pubs.rsna.org/doi/abs/10.1148/radiology.219.2.r01ma19316. [Accessed: 21- Dec- 2017].

[16] G. van Dam, G. Themelis, L. Crane, N. Harlaar, R. Pleijhuis, W. Kelder, A. Sarantopoulos, J. de Jong, H. Arts, A. van der Zee, J. Bart, P. Low and V. Ntziachristos, "Intraoperative tumor-specific fluorescence imaging in ovarian cancer by folate receptor-α targeting: first in-human results", 2017.

[17] R. Israeli, W. Miller, S. Su, C. Powell, W. Fair, D. Samadi, R. Huryk, A. DeBlasio, E. Edwards, G. Wise and W. Heston, "Sensitive Nested Reverse Transcription Polymerase Chain Reaction Detection of Circulating Prostatic Tumor Cells: Comparison of Prostate-specific Membrane Antigen and Prostate-specific Antigen-based Assays", *Cancer Research*, 2017. [Online]. Available: http://cancerres.aacrjournals.org/content/54/24/6306.short. [Accessed: 21- Dec- 2017].

[18] Tasch, M. Gong, M. Sadelain and W. Heston, "A Unique Folate Hydrolase, Prostate-Specific Membrane Antigen (PSMA): A Target For Immunotherapy?", 2017. [Online]. Available: http://www.dl.begellhouse.com/journals/2ff21abf44b19838,46f5c0d0074d488c,6a5b9a2d2ec1da76.html. [Accessed: 21- Dec- 2017].

[19] S. Mannweiler, P. Amersdorfer, S. Trajanoski, J. Terrett, D. King and G. Mehes, "Heterogeneity of Prostate-Specific Membrane Antigen (PSMA) Expression in Prostate Carcinoma with Distant Metastasis", *Pathology & Oncology Research*, vol. 15, no. 2, pp. 167-172, 2008.

[20] J. Trover, M. Beckett and G. Wright, "Detection and characterization of the prostate-specific membrane antigen (PSMA) in tissue extracts and body fluids", *International Journal of Cancer*, vol. 62, no. 5, pp. 552-558, 1995.

[21] J. Ross, C. Sheehan, H. Fisher, R. Kaufman, P. Kaur, K. Gray, I. Webb, G. Gray, R. Mosher and B. Kallakury, "Correlation of Primary Tumor Prostate-Specific Membrane Antigen Expression with Disease Recurrence in Prostate Cancer", *Clinical Cancer Research*, 2017. [Online]. Available: http://clincancerres.aacrjournals.org/content/9/17/6357.short. [Accessed: 21- Dec- 2017].

[22] X. Wang, L. Yin, P. Rao, R. Stein, K. Harsch, Z. Lee and W. Heston, "Targeted treatment of prostate cancer", 2017. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jcb.21491/abstract;jsessionid=0775C19EECCDB611802293FBCA1F86C5.f01t03?deniedAccessCustomisedMessage=&userIsAuthenticated=false. [Accessed: 21- Dec- 2017].

[23] X. Wang, S. Huang, W. Heston, H. Guo, B. Wang and J. Basilion, "Development of Targeted Near-Infrared Imaging Agents for Prostate Cancer", 2017. [Online]. Available: http://mct.aacrjournals.org/content/13/11/2595.short. [Accessed: 21- Dec- 2017].

[24] X. Liora, R. Reddy, B. Matesic, R. Bhargava, "Towards better human capability in diagnosing prostate cancer using infrared spectroscopic imaging", In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. ACM July7-11, 2007; pp. 2098-105. Retrieved from, http://dl.acm.org/citation.cfm?id=1277366

[25] X. Yang, Q. Song, Y. Wang, A. Cao and Y. Wu, "A Modified Deterministic Annealing Algorithm for Robust Image Segmentation", *Journal of Mathematical Imaging and Vision*, vol. 30, no. 3, pp. 308-324, 2008. 1

[26] R. Sammouda, J. Hassan and M. Sammouda, "CT images analysis for early detection of lung cancer", *Int. Innov. Comput. Inform. And Control* (IJICIC), vol. 4, no. 11, pp. 2847-2860, 2008.

[27] R. Sammouda, N. Adgaba, A. Touir and A. Al-Ghamdi, "Agriculture satellite image segmentation using a modified artificial Hopfield neural network", *Computers in Human Behavior*, vol. 30, pp. 436-441, 2014.

[28] A. Petrovsky, E. Schellenberger, L. Josephson, R. Weissleder and A. Bogdanov, "Near-Infrared Fluorescent Imaging of Tumor Apoptosis", *Cancer Research*, 2017. [Online]. Available:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

http://cancerres.aacrjournals.org/content/63/8/1936.short. [Accessed: 21-Dec- 2017].

[29] H. Akbari, L. Halig, D. Schuster, A. Osunkoya, V. Master, P. Nieh, G. Chen and B. Fei, "Hyperspectral imaging and quantitative analysis for prostate cancer detection", *Journal of Biomedical Optics*, vol. 17, no. 7, p. 0760051, 2012.

[30] V. Kondepati, H. Heise and J. Backhaus, "Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy", *Analytical and Bioanalytical Chemistry*, vol. 390, no. 1, pp. 125-139, 2007J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[31] R. Sammouda, A. Al-Salman, A. Gumaei and N. Tagoug, "An Efficient Image Denoising Method for Wireless Multimedia Sensor Networks Based on DT-CWT", *International Journal of Distributed Sensor Networks*, vol. 11, no. 11, p. 632568, 2015.

[32] A. AlSalman, A. El-Zaart, S. Al-Salman, and A. Gumaei, "A novel approach for Braille images segmentation", in *Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS '12)*, pp. 190–195, IEEE, Tangier, Morocco, May 2012.

[33] A. Gumaei, A. El-Zaart, M. Hussien, and M. Berbar, "Breast segmentation using k-means algorithm with a mixture of gamma distributions", In *Broadband Networks and Fast Internet (RELABIRA), 2012 Symposium on*, pp. 97-102. IEEE, 2012.

[34] A. Gumaei, A. El-Zaart, and H. Mathkour, "An efficient iris segmentation approach", in International Conference on Graphic and Image Processing (ICGIP '11), vol. 8285 of Proceedings of SPIE, Cairo, Egypt, September 2011.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

New Image Processing Techniques Using Elitism Immigrants Multiple Objective of Genetic Algorithms for Disease Detection

Khalil Ibrahm Mohammed Abuzanounneh
IT Department, College of Computer, Qassim University Saudi Arabia
ka.abuzanouneh@qu.edu.sa

Abstract— Image processing and analyzing images in the medical field is very important, this research diagnoses and describes developing of diseases at an earlier stage, a detection of diseases types by using microscopic images of blood samples. Analyzing through images changing is very important, the main objective is completed by analyzing evolutionary computation into its component parts, using elitism immigrants multiple objectives of genetic algorithms (EIMOGAs), artificial intelligence system, evolution methodologies and strategies, evolutionary algorithm. EIOMGAs are the type of Soft Computing a model of machine intelligence to derive its behavior from the processes of evolution in nature [1].

The goal of applying EIOMGAs is to enhance the quality of the images by applying the image converting process segmentation to get the best image quality to be very easy to analyze the images. EIOMGAs are the unbiased estimator for optimization technique, and more effective in image segmentation, and it is the powerful optimization technique especially in a large solution space to implement enhancement process. The powerful of EIOMGAs system in image processing and other fields leads to increase popularity and increasingly in different areas of images processing and analyzing for solving the complex problems. The main task of EIOMGAs is to enhance the quality of the image and get required image recognition to achieve better results, faster processing and implement a specialized system to introduce different approaches based on GAs with image processing to obtain good quality and natural contrast of images [2].The development with comparisons used between the different techniques of representation and fitness analysis, mutation, recombination, and selection, evolutionary computation is shown to be an optimization search tools. All features of microscopic samples images and examines change in geometry, texture, colors and statistical analysis will be applied and implemented in this system.

Index Terms— Elitism Immigrants Multiple Objective, Microarray Image Processing, Data Mining, Digital image processing.

I. Introduction

Image processing is a section of artificial intelligence concerned with the enhancement, and analysis of images performed by a computer, and it has become the most important visualization and interpretation methods in biology and medical fields. It has a development of new and powerful tools for analyzing, detecting, transmitting, storing, and displaying medical images, the medical images is challenging to found the development integrated systems, design, implementation, and successful testing of complex medical systems using in the medical aspect, the analyzing process through images is to collect information, diagnosis diseases, diseases detection, and control and therapy evaluation [3].The segmentation and morphological techniques of Digital image processing (DIP) can be applied for analyzing and diagnosis a lot of medical images diseases such as WBCs, the white blood cells play the main goal in the diagnosis and analysis different diseases, the extracting information is very important for hematologists. The different techniques in an image processing are used to analyze the cells to be more accurate and diagnosis systems for remote diseases. There are some complications to extract some data from RBCs in the cells wide variety in shape, edge, position, and size. Moreover, when the illumination is imbalanced, the image contrast between cell boundaries and the background varies based on the capturing process conditions [4].

In the last few years, the image processing techniques got rapidly grown, where hematologists can be used images segmentations of blood automatically, blood slides and blood boundaries for detecting diseases in the diagnosis system.

The research study is focusing on RBCs segmentation process for human blood system using elitism immigrants multiple-objective of Genetic Algorithms and digital images processing. The main goal is to analyze RBCs using EIMOGAs that has been developed in the last years. The using of EIMOGAs in the segmentation techniques of the digital image processing can be applied set of constraints to finding data about the ratio cytoplasm to classify and identify various types of cells such as a lymphocyte, basophil, and neutrophil. The segmentation methods have been applied in many works and different area of images processing, related to region growing, border detection, edge, watershed clustering, and mathematical morphology and filtering processes.

The author proposed an automatic medical system for the segmentation technique and border identification for whole objects based on image boundary among the images database system that is taken from a blood slides and the original image [5], the using of images processing are used as they are not expensive and do not require complex testing and labs equipment, the system focus on

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Thalassemia disease, Thalassemia features in microscopic images and changes in gene geometry, texture, statistical analysis, and colors contrast of RBCs, therefore, the microarray technology can be applied to get a robust genomic system for studying and analyzing the thousands behavior of genes simultaneously. The images analysis which was obtained from the microarray technology strongly helped in the diagnosis, detection, and treatment of most diseases.

In this research can be developed an automated diagnosis system for analyzing and testing data from microscope images directly and detects diseases cases, for that purpose, the digital image processing performs many operations such as modify image rotation, extracting data from the image, locating genes in the images, and the data mining will be normalized the extracted data and getting the effective genes [6].

I. Genetic Algorithm

Charles Darwin is invented Genetic Algorithm as the natural selection process to take input and calculates an output when a set of solutions can be produced. In the last few years, GAs was created to represent processes in the natural system that is important to evaluate and perform an efficient search in the global domain and to have many optimal solutions and more than that. GA is very effective in the contrast improvement in quality and produces an image based on the natural contrast in different scale levels. GAs are the systematic random search techniques to apply generic methods for solving complex problems and optimization process.In the image process, GA can use less information related to the segmentation problems to be solved than the traditional optimization systems, which almost require the derivative objective functions. The fitness function is based on an individual of images, and additionally, GAs can be used a set of different operators (reproduction, crossover, and mutation) to generate new solutions and use it to get an optimal solution for the new images that may contain new chromosomes [7]. Basically, the new children or chromosomes in Genetic Algorithm are obtained of a combination of features of their parents from original images.

The elitism based immigrants multiple objectives of Genetic Algorithms (EIMOGAs) is a new technique will be used in the image processing to produce a set of newly enhanced pixels of the image to be much better than the original image and contains good features, the Image segmentation will be applied EIMOGAs techniques to enhance and improve image quality for extraction more details about the degraded images. The techniques of image colors have some problems such as colors image enhancement applied in the true colour (RGB), where the colour spaces are not suitable for the human system, and the distribution colours in the images are inappropriate the normal visual limits to human perception[8], one technique is not enough to be suitable for one type of image degradations in the RBCs. EIMOGAs have the ability to select optimal colors and segmentation regions to choose appropriate features of the analysis size and select the heuristic thresholds to solve complex problems [7].

II. Thalassemia diseases

There are some main factors will be used in this research to analyze blood color of RBCs, the cells shape, and the cells number, the experiments diagnosis will be checked whether the required factors are negative or positive results, a lot of diseases can happen to cause the size changing cell, shape cell, and the blood cell color. The researchers can be used blood count analyses, blood images analyze, iron analyzes, and the HPLC analyze to check whether the patients are having thalassemia diseases or not. In this research, we proposed a system that can be applied to diagnosis thalassemia disease based on EIMOGAs techniques, The purpose in this work is to help both patients and doctors and health care regarding the reducing time for pathology, the reducing effort, and more accurate in achieving outputs. In this research will be studied two types of thalassemia disease are alpha thalassemia and beta thalassemia. Thalassemia diseases cause a reduction in the lifespan of red blood cells , the disease is a result of an imperfection in the genes that regulate the haemoglobin formation, which is a core ingredient of the red blood cells ,hence thalassemia is hereditary blood disorder characterized by abnormal haemoglobin production and very common in subtropical and tropical areas, for instance, thalassemia disease was infected 280,000,000 people in 2013, with about 439,000 having a dangerous disease, the most common among Middle Eastern people, African descent, Italian, Greek, and South Asian ,both females and Males have similar disease rates, the resulted in 16,800 deaths in 2016, down from 35 thousand, deaths in 1990, so the blood characteristics should be analyzed to make a good diagnosis.

The automated diagnostics system have been developed, using available rule-based tools to cover a blood broader range related diseases containing anemia various types, the alternative automated diagnostic tools are required, in order to find the diagnostic goal, and the differentiation among thalassemia patients, thalassemia traits, and normal people.

The classification problems of thalassemia patients will be formulated in the pattern recognition problems as input process [9]. The test patterns and samples will be blood-related features that are the red blood cell, characteristics reticulocyte, and blood platelet, that is extracted and used in the blood samples.

In the data mining techniques, the researchers are used different rules and patterns to extract data based on the clustering, summarization, association, and classification using the machine learning techniques to test Beta Thalassemia [10]. There are research studies illustrated the Thalassemia testing indicators as Haemoglobin (Hb) A2, Mean Corpuscular Haemoglobin, and Mean Corpuscular Volume. In the Knowledge research, the principal components analyses research were used to discover β-Thalassemia, there are several algorithms for machine learning are applied in the β-thalassemia

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

classification based on new data set which is different from the other researchers, the classifiers of data mining are applied to differentiate among thalassemia traits in different levels as iron deficiency patients, normal people and the patients with other blood diseases [10].

III. IMAGE PROCESSING AND SEGMENTATION

The image segmentation is the partitioning operation of an image into a collection of pixels connected sets, and it also the most significant task in image processing, and for better analysis and diagnosis, the original image will be partitioned into different sizes and pieces. The most important task in the image segmentation is to explore the appropriate parameter selection based on GAs.The purpose of image segmentation are:

1. The regions segmentation to cover the image coordinates.
2. The linear structures segmentations that including line segments and curve segments.
3. The 2D shapes segmentations, such as ellipses, circles, and strips (regions, long, symmetric) for instance, the cluster pixels inside salient image boundaries, the regions corresponding to objects surfaces, or objects natural parts.

The applications of image segmentation include: the image recognition segmentation is using for face recognition, the medical image segmentation such as diagnosis operations, locating cancer diseases and other dangers pathologies. The image segmentation process was used in the agricultural imaging for crop diseases detection. Traffic control system was used to identify shape and size of objects, and moreover, it used to identify moving scene objects using video compression system, the Image segmentation has been divided into two parts of approaches: the region based approach and boundary based approach, in the first part, the purpose is to determine if a pixel belongs to an object or not[11], in the second part the goal is to locate the boundary curves between the background and the objects.

There are four different types of image segmentation:

a) Segmentation greyscale. b) Segmentation texture.
c) Segmentation motion. d) Segmentation depth.

The Main algorithms of region segmentation are divided into three categories:

1. Region-based segmentation technique:

Thresholding method can be used as a simple technique to segment an image for the objects separating from the background using a pixels features values that are compared with a threshold values in order to determine the class of the pixels, this method starts with the first one pixel of a potential region and expands it by inserting adjacent pixels for any image includes different regions, the image will be segmented based on the different areas of the image which each piece has a range of features values, the thresholds are significant to select these thresholds, and it very effectively and useful in the segmentation quality of the images, finally, the statistical test processes used to take a decision which pixels will be inserted into a region segmentation or not.

2. Clustering-based image segmentation technique is dividing the image into different classes which do not require prior information. In the same type of classes, the data should be collected together in similar classes and the data which contains a different type of classes will be in different classes as possible.
3. Edge-based image segmentation Technique is the main features of the original image, which include valuable data useful in image analysis and diagnosis of object classification and explores the detection of boundaries among the different region's image [12]. The boundaries discontinuities occur among the pixels of the selected features such as intensities, textures, and colors.

IV. IMAGE SEGMENTATION USING GENETIC ALGORITHM.

The parameter selection will be applied using EIMOGAs to enhance the parameters selection of the images segmentation and to improve its outputs. The pixel scale and level of segmentation implement GAs will be used to complete region labeling tasks of the image segmentation processes, the proposed method should be used the image adaptive segmentation including the following steps [10]:

1. Compute the image statistics tables give us the probability for a given degrees of a confidence level and identically distributed normally to select suitable threshold. 2. Generate an initial population of segmentation image. 3. The image Segmentation applied initial parameters selections. 4. Compute the segmentation based on quality measures to satisfy conditions of the fitness function.
5. To select new individuals should be used the reproduction operator to generate new population applying by using the mutation and crossover operators.
6. The image segment should be used new parameters to calculate the segmentation quality of an image.
7. Analysis and modify the knowledge based on the knowledge structures of the new image.

V. GENETIC ALGORITHM AND CHROMATIC FEATURES

In this research, we will be applied Elitism Immigrants Multiple Objective of Genetic Algorithms (EIMOGAs) with Chromatic features to describe the color distribution and the grey-level of the images, which are the most discriminative features of Red Blood Cells. the image pixels is represented a segmented object such as (RBCs, RBCs, Nucleus, Cytoplasm, Cells Parasites), The GAs selection operator is used to detect the edge of cells boundaries that have the same colors of pixels from the current population (RBCs images) that will be used new generation.

The convergence process will be completed and achieved in under the iterations required the number to detect RBCs and complete blood counted for a new generation and population.

In the next step of the population, solutions are represented intensity cells colors and chromatic features which can be detected and computed using EIMOGAs of RBCs. In this stage of research will use generation,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

mutation, selection, elitism immigrants, integrate all several immigrants of memory scheme and combined into the EIMOGAs to improve its searching capacity for the image process environment.The image process is a stochastic process where pixels values are modeled as random variables, the GAs can be applied to calculate the probability density of grey level and color distribution as its fitness function [13].

The fitness function is used to get robust convergence as building simulations for RBCs image as possible with reliable convergence and a high convergence with the original image.

the elitism-based immigrants schemes Multi-objective of genetic algorithms (EIMOGAs) efficiently improve the GAs performance in the image processing environment ,and the best selection individual of color pixels based on fitness function from the previous generation is used to create immigrants included probability density of grey level , colors gradient ,colors distribution, cells color and boundaries shapes of RBCs into the population using a genetic operation (selection, evaluation, mutation ,and recombination) , the new process of generation will be the implemented using the elite process e_t from the previous generation g_{t-1} to create new immigrants, as a set of r(e) × n(t) individuals are generated based on fitness function and mutation e_{t-1} with a probability $p(e_{t-1})$, where n(t) is the population number of the image colors , and r(e) is the number ratio of elitism immigrants for each color to the population number, the selection operator of EIOMGAs selects set of cells color of RBCs as the best solutions that have a better classification, based on a fitness functions [14], and then it will be carried forward for recombination image process.

The sensitivity analysis and the results are shown in the final experiments. EIOMGAs are efficiently improve the genetic algorithms performance in the image processing environment, and the best individual from the previous generation (RBCs Image) to next generation can be selected and created immigrants with optimal solutions into the population by evaluation and mutation process.

VI. Proposed System

The proposed analysis system for RBCs segmentation explants the phases are shown in Figure 1. The image pre-processing of the blood smears is applied for removing noises, improving and contrast variation and luminance in the original images. In the second phase, a segmentation processes are applied and implemented to explore and isolate the interest objects of the image. The third phase goal is to extract the objects characters to be used in the next phase of the process, the Features selection method is applied to decrease the redundant data and built classification stage. The selected features are selected for input to the classification method and take the decision about the class assignment by using EIOMGAs as shown in figure 1.

The main goal of the segmentation process is to separate RBCs from another different ingredient of blood image. The blood smear consists four components, the image background, WBCs, RBCs, and cytoplasm. WBC should be darker than the background, and RBCs seem on a high-intensity scale [15]. And also, there are shapes variation in cells and their nucleus.

Figure 1 shows the block diagram of the segmentation scheme.

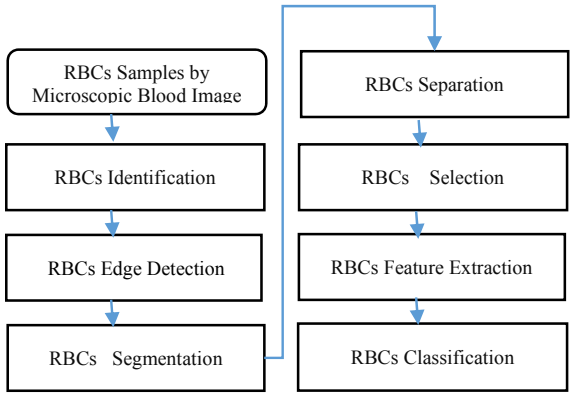


Fig.1.The Proposed Block Diagram of RBCs Analysis and Methods Using EIMOGA

VII. Image Segmentation and Active Contour Models

In the last few years, there are recent developments in the medical imaging fields have brought a new techniques research on image processing for improving medical analysis and diagnosis in segmented images. This technique has been developed to identify specific structures in a magnetic resonance imaging (MRI). The Active Contour methods are adaptable to the desired features in the image.

There are several forms and different types of RBCs images. The applying appropriate method for variable shapes and segmenting for RBCs has been always a challenge for researchers between segmentation methods, the active contour model has a lot of enhancements and implemented in the last few years, In the RBCs, the image should be used active contour models which are changeable curves to respond their change forms to avoid deform objects boundaries in an image segmentation [16].

The active contour models can be moved based on internal or external forces extracted from the image characteristics. The active contour adaptation occurs in response to both internal and external forces, the external forces model has described the gray level gradient, the active contour models can be divided into two types: the parametric models like the Snakes model, which defines a resilient contour that can dynamically adapt to required edges of the image objects, and the geometric models, such as the Level Set model [13] it embeds the front to be zero level set in the higher dimensional function, to calculate the new function evolution, this evolution operation is dependent on the image characteristics extracted and geometric restrictions of the function.

In the processing scheme, the segmentations are implemented on sub-images, the parametric snake model is a curve x(s) defined in Equation.1 [4], to move through

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the image spatial domain and minimize the energy function E(s) defined in Eq.2.

$$v(s) = [x(s) \ , \ y(s)] \ , \quad s \in \{1,0\} \qquad (1)$$

$$E(s) = \int_0^1 \left[\frac{1}{2}(\alpha \mid x'(s) \mid^2) + \beta \left| x''(s) \right|^2 + Ex(x(s)) \right] ds \ (2)$$

Where x'(s) denotes the first derivative, x″(s) denotes the second derivative of x(s). While α, β are parameters of weighting to control the rigidity and tension of snake, respectively. E_{xt} is the function of external energy which is derived from the image to take smaller values of features of boundaries.

The external and internal forces are used the image gradients as a parametric active contour of Snakes models, the external and internal forces will be used the image gradients as a parametric active contour of Snack models. The gradient-based model is better models to use insensitivity to its initial parameters positions and wide capture region of images. Gradient vector flow detriments the object boundaries when addressed locked to the object boundary, while on homogeneous regions will be changed smoothly and will be more extended to the image border, the gradient vector flow field is selected as v(x, y) to be vector field is written in equation.3 [17], which reduces the energy function as defined in the equation. 4.

$$v(x,y) = v(x,y) \ , \ u(x, \ y) \qquad (3)$$

$$E(s_{min}) = \int_0^1 [(\alpha \mid x'(s) \mid^2) + \beta \mid x''(s) \mid^2 + E(x(s))] ds \qquad (4)$$

The internal energy E for the gray-level images I (x,y) is identified as:

$$E_{int} = \int_0^1 \frac{1}{2} [(\alpha \mid x'(s) \mid^2) + \beta \mid x''(s) \mid^2] ds \ (5)$$

While the external energy can be identified as:

$$E_{ext} = -\nabla [(g(x, \ y) * I(x, \ y) \mid^2] \qquad (6)$$

Where g denotes a two dimensional of the Gaussian filter with a normal deviation, ∇ is identified the gradient operator. This filter is applied to the image in order to enhance the map image edge and to reduce an image noise . The regions closer to edges will be given the gradient image high rates. In this research, the cell boundaries will be extracted using edge detection and avoided missing off the edges occurring[18], so the image smoothing using a Gaussian filter to reduce noises with normal deviation is written as the following [17].

$$g(x,y) = G_\delta(x, \ y) * f(x, \ y) \qquad (7)$$

Gaussian's smoothing operator (GSO) is a 2-Dim used to remove noise and detail with special properties as defined in Eq.8.

$$G(x,y) = \frac{1}{2\pi\delta^2} e^{-\frac{x^2+y^2}{2\delta^2}} \qquad (8)$$

There is another problem needed to a solution using additional parameters to improve the external force (k, k1) to improve the capture range in heterogeneous regions of the image edges. This problem will be enhanced using a constant normal (k) [19], to control the external force, the active contour model can be inflated or deflated, based on the sign and magnitude (k1) of the external force. in this paper, the author is proposed applying the balloon model to prevent the snake from stalling in the image homogeneous regions and should be taken to select appropriate values to (k, k1) to make the snake control edges and noise, without exceeding the desired characteristics for contour regions , which is written in Eqau.9.

$$F_{ex} = F_1 \vec{n}(s) - k \frac{E_{ext}}{\|E_{ext}\|}(x(s), y(s)) \qquad (9)$$

Energy Surface and Optimum Thresholding is the basics approach to image segmentation is an amplitude thresholding, a threshold T is chosen to separate the two regions modes, the image point for I(x,y) >T is considered as object points[19], otherwise, the point is called a background point. The threshold method is defined as:

$$g(x,y) = \begin{cases} 0 \ , & I(x,y) < T \\ 1 \ , & I(x,y) \geq T \end{cases} \qquad (10)$$

Where T is set on the entire image basis I(x, y), and the threshold is global. When T depends on spatial coordinates x and y, based on a dynamic threshold, when T depends on both I(x,y) and set property p(x,y) of local image, the average of gray level in a neighborhood centered on I(x,y), the threshold will be local and T is set according to a fitness function is defined by:

$$f(y,x) = \quad T[p(x,y), I(x,y)] \qquad (11)$$

Template Matching is a new type technique in the image segmentation based on prior knowledge of the detected object in image analysis, using the presence detection of an object in a scene, and identifies its position in current given scene [20].

The object locating can be described using a template T[x,y], in the image I[x,y], The best match Searching can be minimized the mean squared errors as written below:

$$E[p,q] = \sum_{x=-\infty} \sum_{y=-\infty} [I[x,y] - T[(x-p),(y-q)]]^2 \ (12)$$

In this research, the correlation technique of images will be used for exploring match of a searched shape w(x, y), of size k*l within an image I(x, y)of a larger size m × n. Where w(x, y) is a search of shape, shape size denotes as z (l, k) , maximum size as m(l, k), the summation will be taken in the image region when w and I not separately, the correlation function techniques have the sensitive disadvantage to local intensities of w(x,y) and I(x,y), the correlation coefficient C(s,t) can be used to remove difficulty pattern matching of local intensities as follows.

$$C(s,t) = \frac{\sum_x \sum_y [I(x,y) - \bar{I}(x,y)] \ [w(x-s,y-t) - \bar{w}]}{\sqrt{\{\sum_x \sum_y \ [I(x,y) - \bar{I}(x,y)]^2 \ \sum_x \sum_y [w(x-s,y-t) - \bar{w})]^2\}}} \qquad (13)$$

In the image, analysis can be used Hough transform technology as a technique of a feature extraction image for RBCs number and to get the number of red blood cell count in the image. Then using machine learning algorithms tool, which has developed a formula to convert a number of red blood cells in the image to actual count by Hough Transform, blood count calculates the blood cells number in a cubic millimeter of blood volume. In this research can be calculated the number of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

RBCs per cubic millimeter based on the cells number in the given image [20].

RBCs count per_cumm = RBCs / ((S /M^2*) * dilution_factor.

Where (cu, mm) is cubic millimeter, and (µl, mcl) is microliters, them is magnification.T is a film thickness. S is an image size.

The Circular Hough Transform technique can be used as a measuring tool to calculate the accuracy using the result of Red Blood Cells number compared with manual counting as the following:

$$\text{Accuracy} = \left(\frac{\text{RBCcount}}{\text{ActualCount}} \right) * 100\% \quad (14)$$

The RBCs classification results need to use set of the numerical analyzing as parameters: mean corpuscular volume (MCV), RBCs distribution width, RBCs count, mean corpuscular hemoglobin (MCH), hemoglobin count (HB), and mean corpuscular hemoglobin concentration (MCHC), for identifying the analyzing combinations.

The hemoglobin (HB) is responsible for the red blood cells color, HB Contents can be calculated by measuring the gradient of colors, the threshold Technique an image will partition into two parts: the foreground and the background.

The binary algorithm was used to calculate the values colors in an image. The classification tools used to formulate a formula and calculate the hemoglobin contents[21][22].

There are three parameters can be identified the red cells characteristics including reporting units, formulas, and definition to calculate each parameter as shown below.

MCV is the RBCs average size constituting the sample. One femtoliter is 10-15 L, the adult's interval (80 - 100 fL) is defined by.

$$\text{MCV} = \left(\frac{\text{Hematocrit} * 10}{\text{RBC} * (10^{12}/L)} \right) * 100\% \quad (15)$$

MCH is the hemoglobin average weight in the RBCs, one pictogram is 10 -12 grams, adults interval (26-32 pg) is defined by

$$\text{MCH} = \left(\frac{\text{Hb (g/gL)} * 10}{\text{RBC} * (10^{12}/L)} \right) \quad (16)$$

MCHC is the hemoglobin average concentration in the RBCs, adults interval (32 - 36 g/dL) is defined by.

$$\text{MCHC} = \left(\frac{\text{Hb (g/gL)} * 100}{\text{Hematocrit}} \right) * 100 \quad (17)$$

The blood samples need to be diluted, so there are some of the RBCs in whole blood to be accurately counted in a microscope[23]. The dilution factor (*Df*) is provided by the industrialist and is around 200X as shown below in equation 18.

$$\text{Actual RBCs} = \frac{\text{RBCs counted}}{(\frac{Ia}{mf^2}) * \text{Ft}) * \text{Df}} \quad (18)$$

VIII. Results and Discussion

This section will be assessed the performance of the proposed RBCs segmentation system. In our experiment, 22 blood samples images of thalassemia. The Image captured was digitized using a digital video camera of Sony high resolution, which was coupled to a Microscope LCD Biological BX5.The experiments were implemented using the Java Genetic Algorithms Package In this research study, there are many methods to diagnosis the Thalassemia diseases. There are three types of images using proposed method to discover abnormal cell types, the color red of blood cells and classification tools.

The research results are described in the following sections:

1- The researcher has identified the abnormal cell types using the various images of red blood cells. The researcher experiments have classified the red blood cells in different shapes. Figure.1 shows the different shapes of red blood cells and the changes in the colors rates (Red, Green, and Blue) of the images after the calculations process have done, the analysis explained the relationship between Thalassemia and non-thalassemia blood images that it has taken as shown in Fig 1.

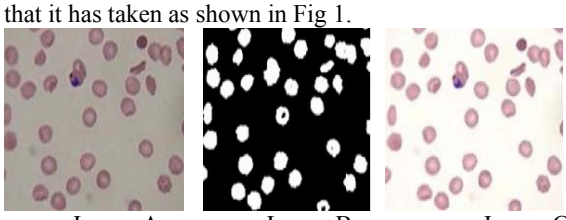


Image A Image B Image C

Fig.1. shows the results of thalassemia infected.

In these experiments , there are different results of Image processing for thalassemia infected, image A is the original image, Image B explained blood image, and Image C shows results after applying pixel classification using EIMOGAs as shown in figure1.

The target cell of blood image, which has hypochromic microcytic and abnormal cell such as sickle cell, that patient will be an affected person of a thalassemia, and for further classification, the author uses other extracted information from classification tools.

2. The RBCs calculations are made to extract the color information (Red, Green, and Blue values) for each image using image processing. In this research, the studies have calculated the average intensity for each of (red, green, blue color), that gave the average rate of colors (red, green, blue) for each sample blood image. The results of average intensity identified the R, G and B rates of thalassemia normal image rate >185 and thalassemia abnormal blood image rate <= 185.

The blood samples need to be diluted, so there are some of the RBCs in whole blood to be accurately counted in a microscope. The dilution factor (*Df*) is provided by the industrialist and is around 200X as shown below in equation 18. The results are reported as the number of RBCs per cubic millimeter of blood.

Males Normal Values = 4.2 - 5.4 million RBCs / mm 3, Females = 3.6 - 5.0 million RBCs / mm 3.

Stained thin blood should be taken by Digital microscope to be more easily distinguished between platelets, RBCs, and WBCs.

To differentiate between RBCs, WBCs, and Platelets, RBCs is less stained as compare to WBC and platelets leaving a bright spot and its intensity value similar the background value.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1. Standard complete blood count System for healthy people.

Blood cell type	Women	Men
RBCs	4-5 M/µL	4.5 - 6.0 M/µL
WBCs	4.5 - 11 K/µL	4.5 – 11 K/µL
Platelets	150 – 450 K/µL	150 - 450 K/µL
Haematocrit	36% to 45%	42% to 50%
Haemoglobin	12 - 15 gm/100 ml	14 - 17 gm/100 ml
Gm /100ml: gram per 100 milliliters; ml: milliliter; gm: grams; µ: Microliter; K: Thousand; M: Million.		

After isolating process of RBCs, we need to apply a counter for counting the RBC number in the image process, so we have used a formula to calculate RBCs number per cumm on the cells number in the given image area (*Ia*) of the blood samples and the of the blood sample film thickness (Ft) is 0.1 mm that is the standard medical system. The magnification factor (Mf) which is the magnification level under the microscope.

In the three experiments below, the assessing RBC morphology procedure includes the smear examination in the thinner edge where the erythrocytes (RBCs) are randomly distributed, the most part singly, with sometimes overlapping cells.

In the next experiment, figure 2 shows three images, Image A is shown an acceptable area of RBC morphology evaluation, we can see that the most cells clearly can be distinguished with some overlapping cells. In an image B is included area too thin, the RBCs appear very flat such as shape cobblestone. An Image C is shown the examined area is thicker than Image B, so the cells will close together, the evaluation process of the morphology will be used individual cells.

Image A **Image B** **Image C**

Fig 2. Segmentation thinner edge area.

The experiments results of the smear examination in the thinner edge area and Assessing Erythrocytes (RBCs) Morphology as shown in figure 2.

Table 2. The Result Comparison between Counted Manually System and CHT using EIMOGA in an Image Processing.

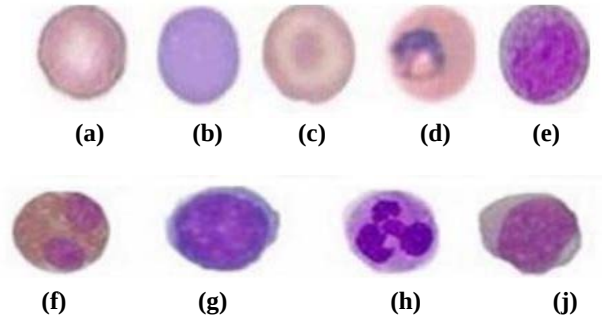
Images	Radius Range In pixels	Film Thickness	RBCs counted Manually System	RBCs counted proposed System
Img1	5-12	0.1	3.768	3.542
Img2	5-14	0.1	4.349	4.161
Img3	4-15	0.1	6.821	6.742
Img4	4-14	0.1	5.783	5.621
Img5	5-13	0.1	4.981	4.813
Where: Magnification Factor(Mf)=300*300; Dilution Factor(Df) =200 ;				

Table 1. Type Sizes for Camera-Ready Papers

After the calculations that have done. In Fig 5 shows the changes in the colors rates (Red, Green, and Blue) of the two Images.The analysis explained the relationship between Thalassemia and non-thalassemia blood images that it has taken.

In the research observations, based on the research details and results, could able to analysis possible diseases combinations when high or low above according to values of parameters.

In figure 3, shows if there is a normal or abnormal of MCV, MCH, RBC and MCHC attributes of a patient's blood sample, and this analysis can be diagnosed the current disease for that patient.



(a) **(b)** **(c)** **(d)** **(e)**

(f) **(g)** **(h)** **(j)**

Fig 3. Different Images of RBCs samples using EIMOGA

Figure.3 shows different Images of RBCs samples using EIMOGA as the following: the images (a) and (b) are healthy RBCs. the images (c) and (d) are infected RBCs with ring parasites. The Images (e) and (f) are infected with trophozoites parasites. The images (g), (h) (j) are RBCs infected with schizont parasites.

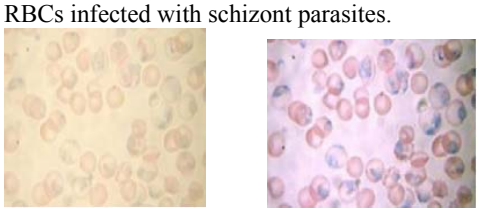


Image A **Image B**

Fig4. Compression between original and enhanced image.

In figure 4, Image A corresponds to original Image, and Image B displays enhanced image after applying EIMOGAs approach for image filtering.

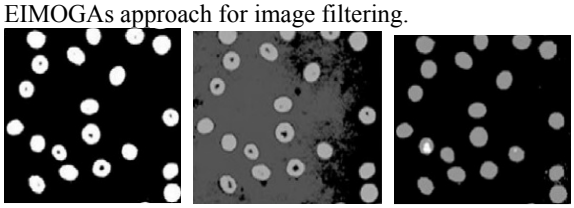


Image C **Image D** **Image E**

Fig 6. Shows different method segmentation results.

Figure 6 shows three different method segmentation results of images. Image C is gotten using EIMOGAs algorithm of the histograms approach. Image D is obtained by applying EIMOGA for pixels classification. Image E is used EIMOGA for clustering approach.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

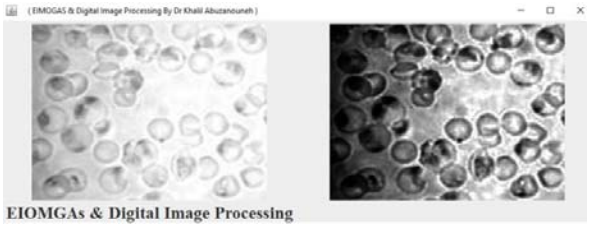


Image A Image B

Fig 7.Shows EIOMGAs technique to enhance Images results.

The experiments results using EIOMGAs technique to enhance the contrast of images,Image A is gotten using EIMOGAs algorithm of the histograms approach. Image B is obtained by applying EIMOGA for pixels classification as shown in figure 7.

In this experiment as shown in table 3, the Original Image number is 5 of microscopic image, the accuracy average is 93.33% was implemented. The comparing process is implemented between the input image and an image after using Circular Hough Transform, there are some of RBCs is not calculated in CHT method due to deformable shape and another condition.

Table 3. Shows the RBC results are counted for 5 RBCs Images using CHT method.

Original Image	Actual RBCs	CHT method Counted	Accuracy of CHT
Img_RBCs_1	903	837	92.69
Img_RBCs_2	937	871	92.96
Img_RBCs_3	978	916	93.66
Img_RBCs_4	1018	957	94.01
Img_RBCs_5	1123	1048	93.32
			Average= 93.33

In the next experiment, we have the same number of Original Images from the previous experiment microscopic image, the accuracy average is 97.05 was implemented. The comparing process is implemented between the input image and an image after using EIMOGA to enhance analysis process, the final results were more accuracy for the counted process of RBCs by EIOMGA as shown in table 4.

Table 4. Shows the Accuracy Results for 5 RBCs Images using EIMOGAs method.

Original Image	Actual RBCs	EIMOGAs Counted RBCs	Accuracy of EIMOGA
Img_RBCs_1	903	877	97.12
Img_RBCs_2	937	913	97.44
Img_RBCs_3	978	949	97.03
Img_RBCs_4	1018	987	96.95
Img_RBCs_5	1123	1086	96.71
			Average= 97.05

IX. REFERENCES

[1] Khalil Ibrahim Mohammad Abuzanouneh. “Hybrid Multi Objectives Genetic Algorithms and Immigrants Scheme for Dynamic Routing Problems in Mobile Networks”. International Journal of Computer Applications 164(5): 49 -57, April 2017.

[2] Khalil Ibrahim Mohammad Abuzanouneh “Parallel and Distributed Genetic Algorithm Multiple Objectives to Improve and Develop of Evolutionary Algorithm, International Journal of Advanced Computer Science Applications, Volume7 Issue 5, 2016.

[3] Ritter Nand CooperJ(2007)Segmentation and border identification of cells in images of peripheral blood smear slides. In Proceedings of the Thirtieth Australasian Conference on Computer Science 62:161–169.

[4] Gonzalez RC, Woods RE, Eddins SL (2003) Digital image processing using MATLAB. Prentice-Hall, Upper Saddle River.

[5] Ritter Nand CooperJ(2007)Segmentation and border identification of cells in images of peripheral blood smear slides. In Proceedings of the Thirtieth Australasian Conference on Computer Science 62:161–169.

[6]

[7] Cseke I (1992) A fast segmentation scheme for white blood cell images. In Pattern Recognition. Conference C: Image, Speech and Signal Analysis, Proceedings. 11th IAPR International Conference on vol. III:530–533.

[8] SobelI(1978) Neighborhood coding of binary images for fast contour following and general array binary processing. Comput Graph Image Process 8:127–135.

[9] Prewitt JMS (1970) Object enhancement and extraction. Picture Processing and Psychopictorics, pp. 75–149.

[10] NHLBI. "How Can Thalassemias Be Prevented?". 3 July 2012. Archived from the original on 16 September 2016. Retrieved 5 September 2016.

[11] ZackG, RogersW, Latt S(1977)Automatic measurement of sister chromatid exchange frequency. J Histochem Cytochem 25:741–753.

[12] Otsu N (1975) A threshold selection method from gray-level histograms. Automatica 11:285–296.

[13] Ongun G, Halici U, Leblebicioglu K, Atalay V, Beksac M and Beksac S (2001) Feature extraction and classification of blood cells for an automated differential blood count system. Neural Networks. Proceedings. IJCNN '01. International Joint Conference on vol. 4:2461–2466.

[14] Kan Jiang, Qing-Min Liao and Sheng-Yang Dai (2003) A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. Machine Learning and Cybernetics, 2003 International Conference on, vol. 5:2820–2825.

[15] Dorini LB, Minetto R and Leite NJ (2007) White blood cell segmentation using morphological operators and scale-space analysis. In SIBGRAPI '07: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing.

[16] Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Comput Graph Image Process 1:321–331

[17] Canny J (1986) A computational approach to edge detection. IEEE Trans. Pattern Anal Mach Intell, pp. 679–698.
[18] Xu C, Prince JL (1998) Snakes, shapes, and gradient vector flow. IEEE Trans Image Process 7:359–369.
[19] Xu C and Prince J (1997) Gradient vector flow: a new external force for snakes. In 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings, pp. 66–71.
[20] Kumar BR, Joseph DK, and Sreenivasc TV (2002) Teager energy based blood cell segmentation. Digital Signal Processing. DSP 2002 14th International Conference on, vol. 2:619–622.

Dr. Khalil Ibrahim Mohammed Abuzanouneh is assistant professor in Computer Collage at the Department of Information Technology, Qassim University. In teaching, he has been focusing on applying genetic algorithms and solves problem based on genetics approaches in Computer Education. In research, his current interests include optimization problem and data compression and security. Dr. Khalil received his PhD degree in Computer Engineering Ph.D.in Computer Engineering Sciences “Computer Aided Design”.
.

Reliability of Scalable Coherent Interface in Cluster Interconnection

Mr Adinew Belay[1], Dr Narasimha Rao Yamarthi[2]

HOD, Department of CS, SCI, Mizan Tepi University, Eithiopia.
Professor, Department of CS, SCI, Mizan Tepi University, Eithiopia.
adinewb2@gmail.com[1],narasimha.yamarthi@gmail.com[2]

Abstract: In recent advanced technology parallel computing plays a vital role in High performance computing. The processor interconnection is one of the prominent factor decides the enactment of high performance computing. The clustering of processors in parallel computing is a challenging job for the present engineers as it depends on many parameters to monitor in packet communication. In the present paper awell-organized scalable coherent interconnection is presented to improve the efficiency of comprehensive parallel computing and cluster reliability. The reliability analysis of redundant fault-tolerance systems is discussed with Markov Modelling. This can be greatly helps to design an efficient and fault tolerance cluster interfacings. The Symmetric clustered processors are designed using Proteus simulation tool.
Keywords: Cluster Interconnection, Markov model, Coherent Interface, Reliability, Scalable system

1. Introduction

Multiple processors used in a computer system enhance the speed of the user operations. Adding second processor is very easy and economical when compared with the second additional computer. Multiple CPUs are interconnected to meet the above purpose. Multiple CPUs are used to perform multiple tasks in single time. Each CPU dedicated with individual task. Such as one CPU controls the operating system programs, and the other CPU may control either memory or I/O operation. Multiple programs with multiple set of instructions can be executed by individual processor [1]. Multiple CPUs are connected in single computer and they are able to perform multiple operations simultaneously and can simultaneously allocate tasks to individual processes or programs [2][3][4]. Although there are multiple CPUs, there are still some difficulties need to be addressed for reliable operation in the computer system [5].

Multiple processor interconnections are implemented such that each CPU and each interconnection behaves separately with or with-out matching their functionalities. There are few limitations need to be face by adding multiple CPUs or second CPU in the multiprocessor system. In general the Operating System (OS) developed in a system for supporting and configuring single CPU. On the other hand when second CPU is connected, then operating system took log time to configure the second CPU and also takes long time to run the tasks in the second CPU. Hence to handle several CPUs it took long time to make then perform reliable and efficiently. This problem is overcome by extend its minimal services of operating system with the second CPU too. The extending time period help to configure the second time. But, the extended time periods slow down the all tasks, and make it enter into boot process. On the other hand, the OS executes the tasks given at the primary CPU. The alternate process extending the services of the OS may not satisfy the actual requirements of the parallel processing. Indeed it is decreasing the performance and led to diminishing the reliability. On the other hand the common OS allowed to run all processors, but not allowed to run programs on peripherals/IO devices or on particular peripheral or on particular peripheral on particular CPU [6].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In multiprocessing each processing system shares a common main memory and peripherals for simultaneous allocation of tasks [7]. It is not always true that multiple processor uses single task or process simultaneously. The systems which share all CPUs in the same way are called Symmetric Multi-Processing (SMP). The Systems which does not treat all CPUs in the same way and the resources are used in different ways are called Asymmetric Multi-Processing (ASMP), clustering multiprocessing, and Non-Uniform Memory Access (NUMA). In the present work symmetric multiprocessing is implemented using Proteus simulation tool.

The processors are either tightly or loosely coupled in multiprocessor systems. In tightly coupled systems the CPUs are coupled at the bus level and shared with common central memory or in some systems hierarchical memory system is shared. The size of the tightly couple processor is physically small and perform is better than loosely coupled multiprocessor system. The chip type multiprocessor systems and multiprocessors in mainframes are of type tightly coupled. The standalone and single or dual processors are connected through high speed Ethernet in loosely coupled multiple processor systems. Loosely coupled multiprocessors are less expensive when compared with tightly coupled multiprocessor system. Tightly couple systems are efficient in power consumption. Loosely coupled systems can work at different operating systems and versions.

2. Computer Cluster

The computer cluster may contain either tightly coupled or loosely coupled multiprocessor system. The coupled systems in multiprocessor system work together to form a network or to form a single system. Each node in the computer cluster is controlled, scheduled, and managing the tasks by software [8].Each node in the cluster is connected with local area networks. In most of the instances the nodes in the cluster use same type of hardware and OS. In few applications like open source applications different hardware and operating systems are used in cluster nodes [9][10]. The load in the computing system will be distributed to all nodes in the cluster. This will optimize the load queries assigned to the cluster nodes. Distributing of load to optimize the load is called load balancing. The cluster computing is used in simulated weather analysis rather than database analysis. The cluster performance mainly depends on scalability, low maintenance, centralized management, and resource allocation.

3. Cluster Management

One of the major challenges in cluster is management of all processors in the cluster. Sharing of memory in cluster is difficult to manage and cost effective. In a heterogeneous processors cluster the individual tasks are given and the performance of the job is decided by the models and characteristics of the cluster. Hence mapping of various tasks into clusters produce major challenges. When a node in the cluster the method called fencing plays an important role in making the system operative. The fencing can be done in two ways. One is deactivate the particular failed node or disallow others to access the resources [11]. The CPUs in the cluster randomly changes to estimate the future states in the operations allocated to them. The Markov model is the stochastic method used to design randomly changing system, which are depending on present state but not on past state.The Markov random field depends on the neighbouring states in multiple directions. Based on the observations made on system, the systems are divided into two methods. Neighbouring states can be estimated by distributing the random variables. The distribution further depends on neighbouring variables. One is autonomous system, which includes Markov chain model, Hidden Markov model. The second is controlled

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

system which includes Markov decision process and partially observable Markov decision process [12][13].

3.1 Markov Chain: It is a simplest Markov model. Markov model which depends on previous state. It states that the randomvariables vary as the time passes.

3.2 Hidden Markov Model: It is markov model in which states are only partially observable. But those observations are not sufficient to define the state. These sequence of observations are evaluated by forward algorithm and starting probabilities, observation functions, are evaluated by Baum-welch algorithm.

3.3 Markovdecision Process: The transitions in the model depend on current state and the action vector. Reinforcing algorithms are implemented in this process and solved with iteration methods.

3.4 Partially observable Markov decision process: It is a process in which states are only evaluated partially. They are mostly used in artificial intelligence applications such as robotics.

4. Architectural Design

In the present work Symmetric Multiprocessors are interconnected and data transfer is observed using simulation tool Proteus. In symmetric multiprocessor system two or more processors are connected to common system bus and shared a common main memory. All the processors are grant permission to access memory and IO devices. A single OS is used to control all processors and IO systems. The polling of IO devices are controlled by interrupt controller programs. The SMP architectural block diagram is shown in figure 1.

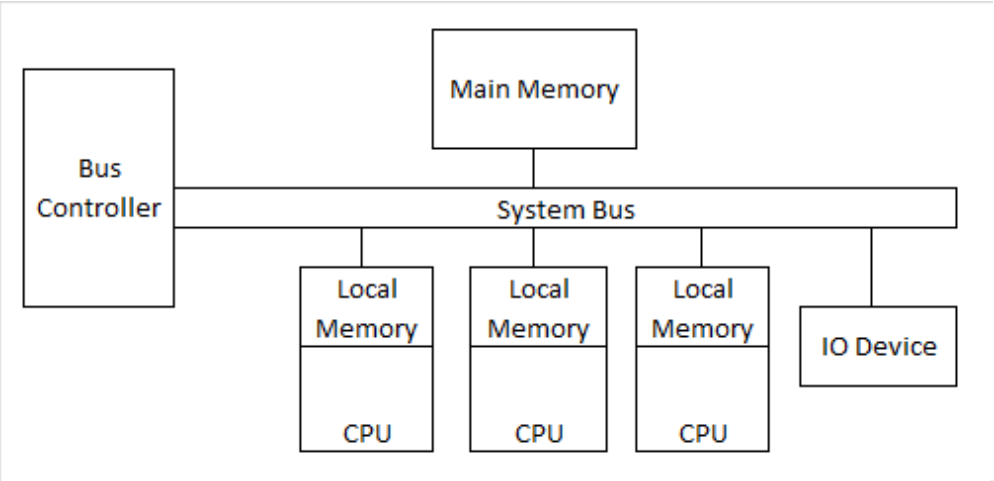


Figure 1: Architecture of Symmetric Multiprocessor System

The bus controller will control the direction of data and produces the control signals like read, write, and interrupt. Memory and IO selections are done by bus controller. All processors have equal priority to access the main memory. The fetching from the memory is controlled through daisy chain method. The memory access is done through system bus. Only one operation either IO transfer or Memory transfer is done with the help of bus controller.

In the present architecture Markov network is considered for analysing the processor operations in multiple dimensions. Unlike Markov chain the Markov network each state depends on neighbours located at different directions. Markov network is visualised like graph of variables.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Here the reliability of the cluster-based system is defined as follows,

$$R_{cl}(t) = R_c(t)\sum_{l=2}^{k} p_l(t) \text{ ---------- (1)}$$

Where,R_c (t) is the reliability of the bus controller, and the probability of *l* number of connected functional clusters with the system at time t is $p_l(t)$is given by,

$$p_l(t) = C_C^{k-1}\binom{k}{l}(R'_C(t))^l(1 - R'_C(t))^{k-1} \text{ ---------- (2)}$$

5. Results

In the current work the throughput is observed in data transfer from cluster to memory. The throughput is observed while all processors in clusters are interconnected without any disconnection between them. The interconnection is framed so that the data successfully transferred to destination even if any processor node failed or disconnected. When any node is failed the data transfer from the failed node to be transferred will be bypassed and the data will be bypassed to next or neighbouring node. Then the data of failed will be passed to destination through neighbouring node along with the new bypassed node data. The table 1 shows the number of active nodes and their respective throughput.

Table 1: Throughput Vs failure nodes

Number of failed Nodes	Throughput (µs)	
	Traditional Model	Markov Model
0	0.825	0.825
1	1.15	1.05
2	1.40	1.25
3	1.65	1.45

6. Conclusion

Hierarchical Markov model is applied in the present work to classify the states of CPUs present in the multiprocessor system. For instance switching the states between memory fetching, IO device controls by multi-processor is observed. Efficient and reliable transfer of data and switching the states between processor is successfully handled. The present analysis is proved that the Markov models are effectively implemented to subvert architectures of advanced cluster-based systems. The reliability is enhanced in terms of throughput and observed in case of failure nodes also.

References

1. Introduction to Multiprocessing (http://www.bitsavers.org/pdf/dec/pdp11/1174/ Introduction _to_Multiprocessing _Jun79.pdf): distinguishes “symmetric” from “master/slave”.
2. Raj Rajagopal, “Introduction to Microsoft Windows NT Cluster Server: Programming andAdministration”,CRC Press, 1999, p. 4.
3. Mike Ebbers; John Kettner; Wayne O'Brien; Bill Ogden, “Introduction to the New Mainframe: z/OS Basics”,2012, IBM. p. 96. ISBN 9780738435343.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

4. Chip multiprocessing (http://www.computerworld.com/hardwaretopics/hardware)
5. Kumar S et al., “Optimizing Technique to Improve the Speed of Data Throughput through ipeline”, Information Systems Design and Intelligent Applications, 8-9 Jan 2016, by SPRINGER Advances in Intelligent Systems and Computing (AISC series), Vol 435.
6. Early Computers at Stanford: the dual processor computer at the AI labhttp:// forum.stanford.edu/wiki/index.php/Early_Computers_at_Stanford#DEC_PDP10_2.
7. Irv Englander, “The architecture of Computer Hardware and Systems Software”, An Information Technology Approach, 2009, 4th ed., Wiley.p. 265.
8. IEEE Technical Committee on Scalable Computing, TCSC, https://www.ieeetcsc.org/.
9. Kumar NS et al., “A New Method to Enhance Performance of Digital Frequency Measurement and Minimize the Clock Skew” in IEEE Sensor Journal, vol 11(6), 2011.
10. Kumar et al., “Design of Intelligent Multinode Sensor Networking”, International Journal on Computer Science and Engineering (IJCSE), vol2 (3), 2010, pp. 468-472.
11. K. Shirahata, et al., “Hybrid Map Task Scheduling for GPUBased Heterogeneous Clusters”, Cloud Computing Technology and Science (CloudCom), 2010 Nov. 30 to 2010 Dec. 3, pp. 733 740.
12. Leslie Pack Kaelbling; Michael L Littman & Anthony R Cassandra, "Planning and acting in partially observable stochastic domains", Artificial Intelligence (Elsevier), 1998, Vol 101, (1–2): pp 99–134.
13. H. H. Bui, S. Venkatesh, and G. West,“Policy recognition in the abstract hidden markov model”, Journal of Artificial Intelligence Research, 2002, vol. 17, pp. 451–499.
14. Yue Luo et al., “Analyzing and Improving Clustering Based Sampling forMicroprocessor Simulation”, Proceedings of the 17th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD’05), 2005, IEEE.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Effort Prediction Tool using UML diagrams (EPT)

[1]Atica M. Altaie, [2]Asma Yassin Hamo

[1] Atica_swe@yahoo.com, [2]Asmahammo@yahoo.com
[1,2] *Software Engineering Department, Computer and Mathematics Science College, University of Mosul, Mosul, Iraq*

***Abstract*-The use case and class diagrams are important models of the system created during early phases of the software development. Effort and size estimation are also important points of the software development. Many effort estimation models proposed in the last years and many factors have an impact on software efforts like complexity, use case points and class points. Effort/size estimation is calculated using the proposed model online shopping system as a case study. The results indicate that the proposed model can help to estimate project size earlier in the design phase, to predict effort needed to complete development. The percentage of estimated effort between two diagrams is 85.49 is obtained .**

***Keywords*:** *Effort estimation, Use case points, Class points , Project size, Project complexity, Metrics.*

I. INTRODUCTION

Software effort is used to measure the use of the workforce. It is the total time that the members of a development team required to perform a given task . It is usually expressed in units such as man-hours , man-day , man-month , man-year. This serves as an indicator for estimating other values relevant , like cost [8].

An accurate estimation of effort is the most important factor in industries [7]. Both under estimation and over estimation can cause severe problems such the underestimation leads to under staffing and consequentially takes longer to deliver project than necessary. Over estimation may lead to miss opportunities to offer funds for other projects in future [9]. To avoid this, human experiments are needed to judge on the results, so the researchers try to develop models for accurate software effort estimation [7].

Line of code is a very important unit for time and effort estimations and many researchers denoted that a count in LOC depends on the degree of code reusing and can be accurate five times higher than anther estimate[5]. So, empirical studies have had an important role in the evaluation of tools, methods before they introduced in real software[5]. The UCP equation is composed of three variables: Unadjusted use case points (UUCP), The technical complexity factor (TCF), The environment complexity factor (ECF). The UCP method is versatile and extensible to a variety of development and testing projects. It is easy to learn and quick to apply [15].

In this paper the impact of UML diagrams on effort estimation is explored. The focus will be on the use case and class diagrams which are pure measures of size and they can establish an average implementation time of project development. It aims to:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- *Propose a method for estimating OO software size, complexity and effort needed for the software to be developed.*
- *Apply the above metrics for a project of on-line shopping, as a case study.*
- *Find the percentage of estimated effort between the two diagrams.*

This paper is organized into 6 Sections. The related work is described in Section 2. The effort prediction design, properties, metrics and tool architecture are described in Section 3. The case studies and results are described in Section 4. Finally, conclusion and future work are described in Section 5 and Section 6 respectively.

II. Related Work

Anda et al [1] estimated the software effort based on use case components and compute the total time in hours. Lavazza and Robiolo [2] showed the measurement-oriented UML modeling can support the computing effort based on functional size and complexity as independent variables. Sridhar [3] proposed knowledge based effort estimation for multimedia projects and concluded that the accuracy of effort estimation can be improved using knowledge rules. Harizi [6] defined parameters of class diagram with their importance, complexity and studied their impact on software size estimation.

Azzeh and Nassif [9] aim to study the potential of using Fuzzy Model Tree to derive effort estimates based on UCP size measure. Bardsiri and Hashemi [10] produced a brief review of well-known approaches from software effort estimation , classified as algorithmic and non-algorithmic techniques, summarized several models with some aspects impacting effort and concluded that each model has its own environment to be effective. Alves et al [5] described a case study based on function points with two teams that developed a software for a real customer to estimate the size and complexity of a software .Saroha and Saha [11] tried to answer questions dealing with factors that impacting effort estimation and to give guidelines for getting accuracy of estimated effort.

Kirmani and Wahid [12] studied 14 projects and applied proposed model in case of use case point and approved its improvements on estimated effort. Kirmani and Wahid also [13] observed scalability in technical complexity factor, project methodology in environmental complexity factor and their impact on estimated effort. Whigham et al [4] proposed a transformed linear model as a suitable baseline model for comparison of software effort estimation methods.

III. The Proposed Model

First, we required to draw the use case and class diagrams for a specific system in enterprise architect tool (Sparx Systems Enterprise Architect, a UML 2.1 based modeling tool for designing and constructing software systems) [16], then generate an XMI file for each diagram and use them as inputs to the effort prediction tool (EPT). Second, will be used a number of special metrics in software engineering to estimating size, complexity and effort of the project through the following steps:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Use case points

1) Computing use case complexity.

1.1) Calculate Unadjusted actor weight (UAW) by summation of a number of actors (NOA) multiplied with their weight.

1.2) Calculate Unadjusted use case weight (UUSW) by summation of a number of use cases (NOUC) multiplied with their weight

1.3) Calculate the number of roles (NOR)

2) Calculate unadjusted use case point (UUCP)

3) Calculate technical complexity factors (TCF) and environmental factors (EF) : We used Seventeen standard technical factors to estimate the impact on productivity and eight factors to estimate the impact on environment [14]. Each factor is weighted according to its perceived impact.

4) Calculate use case points (UCP)

*5) Assume that PF=20, then calculate Effort = UCP * PF .*

B. Class points

1) Computing class complexity.

1.1) Calculate the number of state point (SP) of class through summation the total functions in each Class multiplied by its own weight.

1.2) Calculate the behavioral point (BP) of class through summation the total number of Method in each Class multiplied by its own complexity and the result multiplied by one plus the number of associations per the class.

2) Calculate the number of class points in the project (CP) Size of class.

3) Calculate the size of each class in the class diagram .

4) Calculate the Size of a system .

5) Calculate effort based on the size of the system.

Figure 1 illustrates the steps of execution represented by an activity model for use case points and figure 2 illustrates the steps of execution represented by an activity model for class points. The enterprise architect tool is used to draw the models [16].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

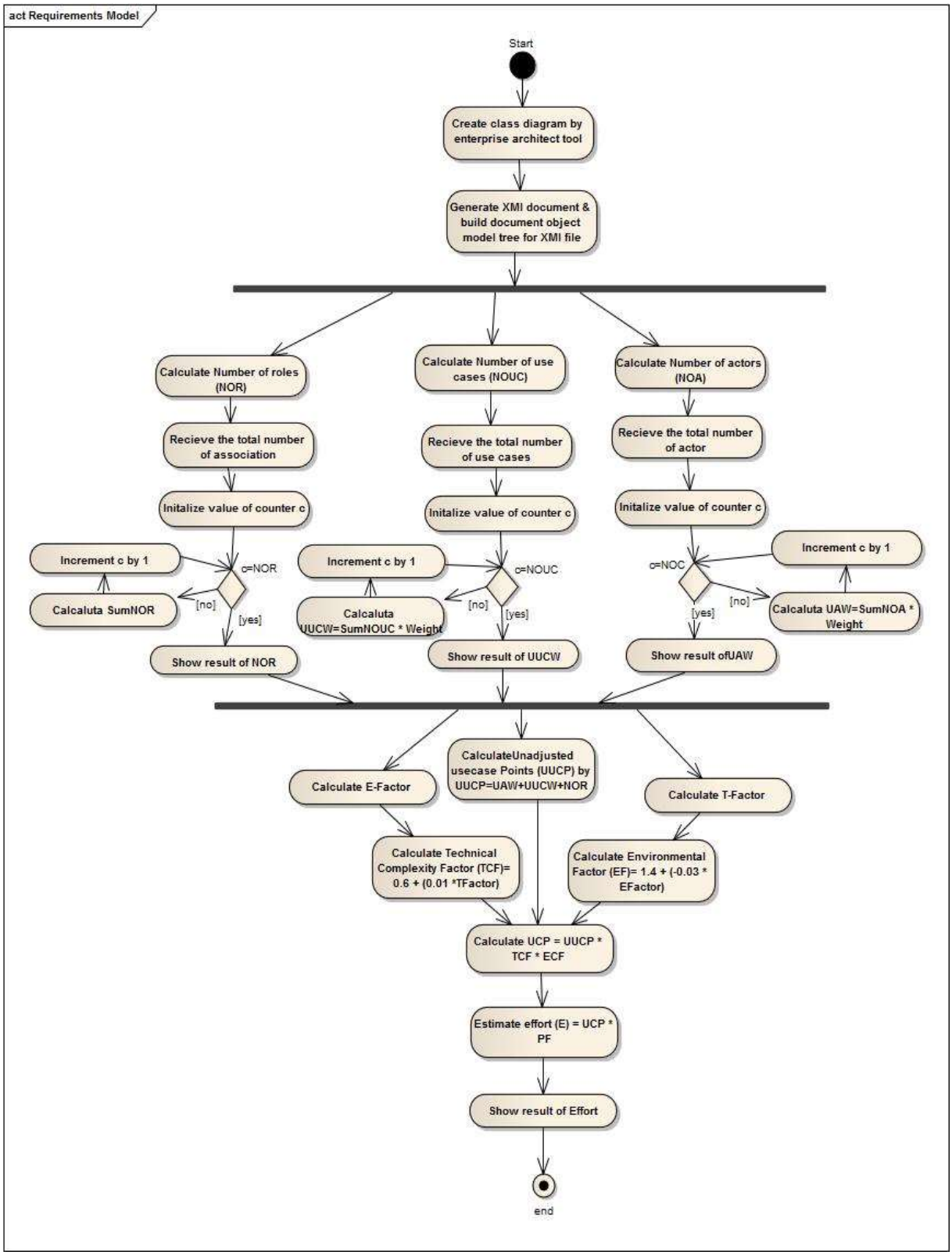


Figure 1. Activity model of use case pionts.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

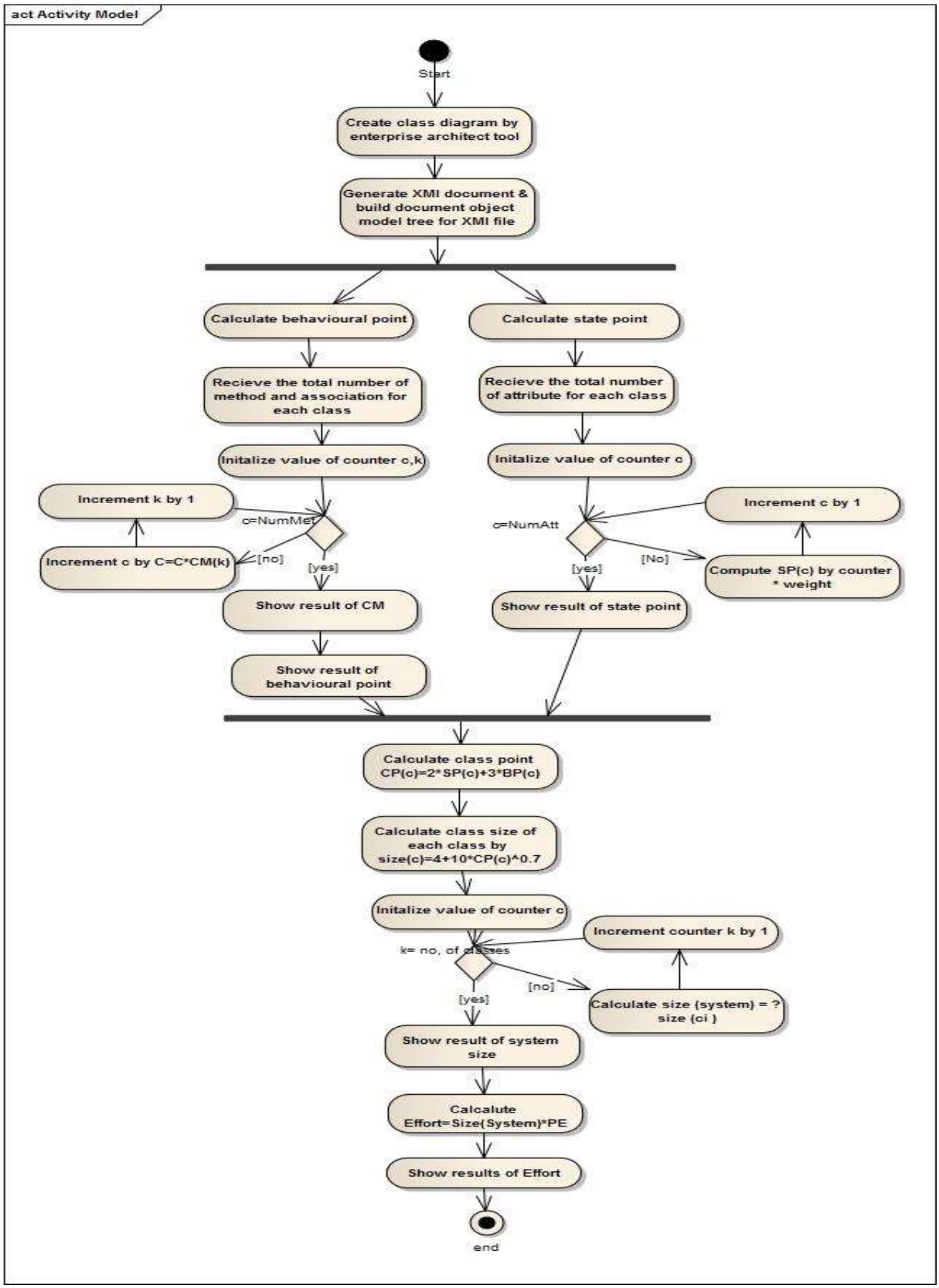


Figure 2. Activity model of class pionts.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. Proposed model Testing and Results

This section explains testing of the proposed model and results that are obtained after executed it.

4.1) Case Study

A case study of an online shopping system is taken and the metrics of the proposed model to obtain the size of project and effort needed. Figure 3,4 illustrate the use case and class diagrams for online shopping system and table1,2 illustrate the detail results obtained respectively.

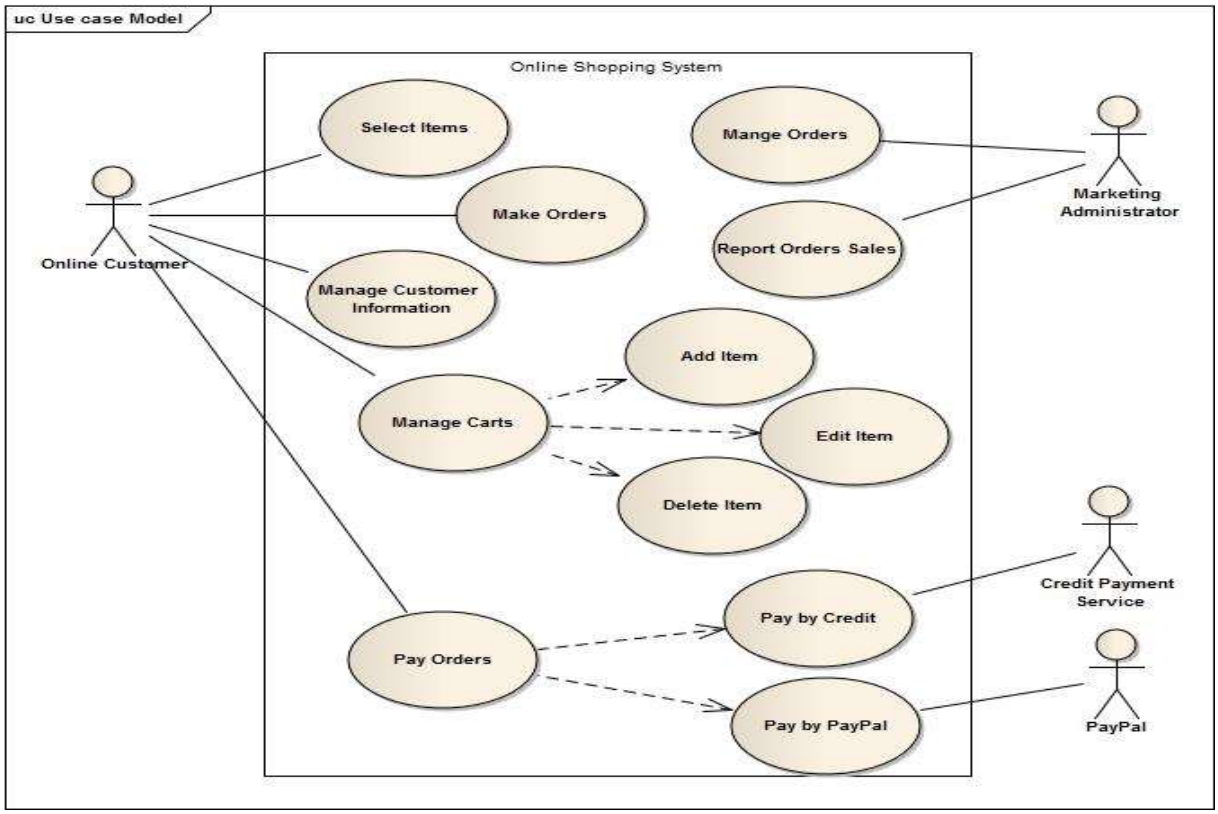


Figure 3. Use case diagram for online shopping system.

TABLE I
Results of use case diagram to predict effort

Description	Variables	Value
Number of actors	numActor	4
Number of use case	numUsecase	7
Number of rules	numRole	14
Unadjusted actor weight	UAW	10
Unadjusted role weight	URW	28
Unadjusted use case point	UUCP	98
Technical Complexity Factor	TCF	1.1
Environmental Complexity Factor	ECF	0.86
Productivity Factor	PF	20
Use case point	UCP	92.7
Effort	E	1854 man/hours

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

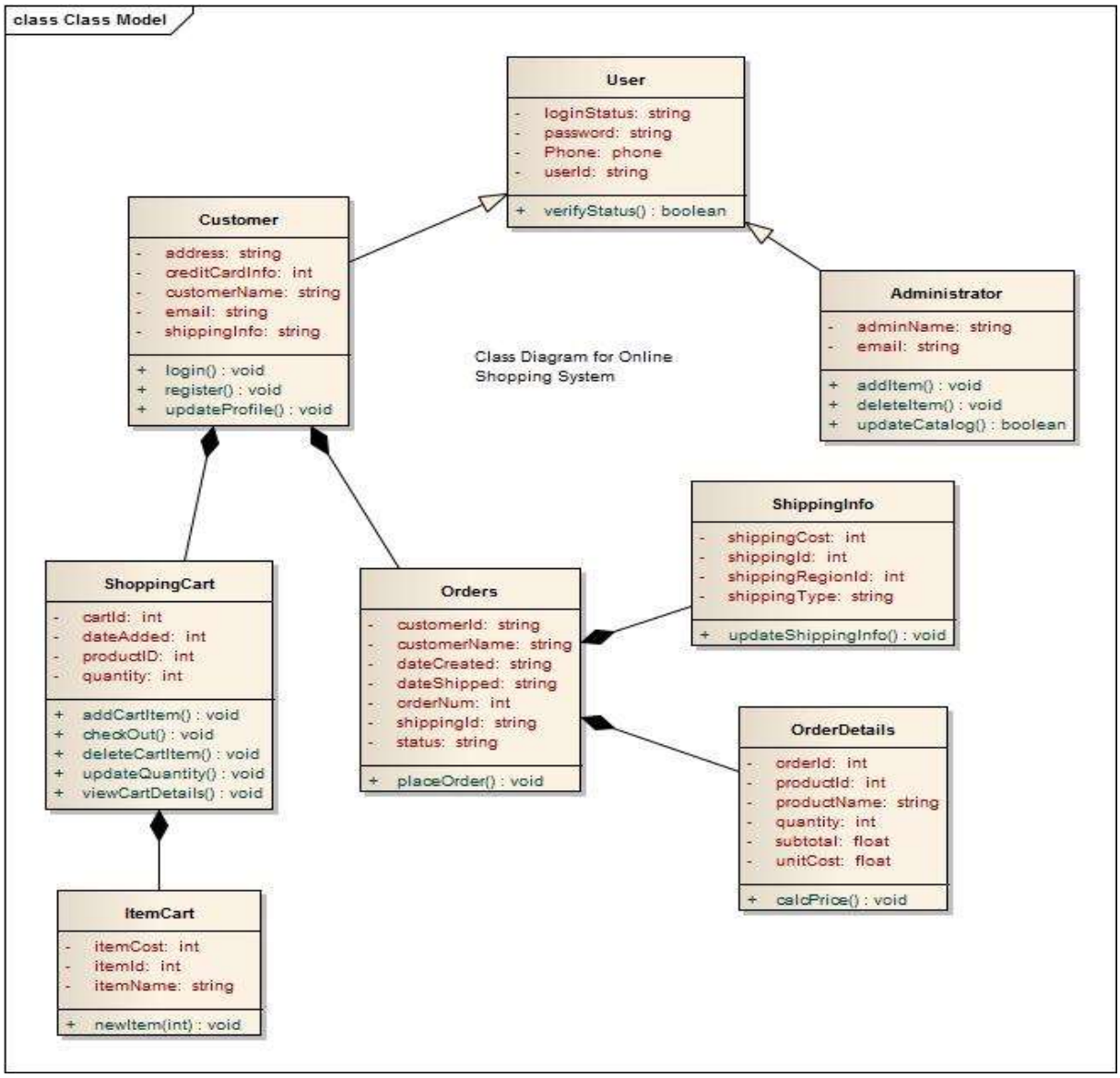


Figure 4. Class diagram for online shopping system.

TABLE II
RESULTS OF CLASS DIAGRAM TO PREDICT EFFORT

Class Name	SP	BP	CP	Size(Class)
User	4	3	17	76.66
Customer	5	12	46	149.86
Administrator	2	6	22	91.04
ShoppingCart	4	15	53	165.06
Orders	7	4	26	101.83
ShippingInfo	4	2	14	67.43
OrderDetails	6	2	18	79.63
ItemCart	3	2	12	60.94
Size(System)=792.45				
Effort=1585 man/hours				

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. CONCLUSION

This research presented a method for estimating OO software project size and the effort needed exploiting UML diagrams. So, through the building and testing of the proposed model, conclusions are:

- *The proposed model metrics can help software engineers to estimate project size and complexity in terms of lines of code earlier in the design phase.*
- *The proposed model can help to predict effort needed to complete development of the project easily in terms of man/hours and to give indicator for managing the overall budgeting and planning.*
- *The percentage of estimated effort between class and use case diagrams is 85.49 .*

VI. FUTURE WORK

In the future, the work may be enhanced in the following aspects.

- *Estimate effort of proposed model can be expanded using information extracted from sequence diagrams, activity and state chart or other diagrams of UML.*
- *The UML points can apply to more projects to provide guidelines for how to measure effort in different kinds of projects.*
- *The proposed model accepts only XMI documents generated by EA, so a model can be extended to accept XML documents also.*

REFERENCES

[1] B.C.D Anda, H.C Benestad, and S.E Hove,"A Multiple-Case Study of Effort Estimation based on Use Case Points", International Symposium on Empirical Software Engineering, ISESE'2005. IEEE Computer Society, Noosa Heads, Qld., Australia, Nov. 17-18, pp. 407–416, 2005.

[2] L. Lavazza,G. Robiolo, "Introducing the evaluation of complexity in functional size measurement: a UML-based approach", in: Proceedings of the 4th International Symposium on Empirical Software Engineering and Measurement – ESEM 2010, Bozen, Italy, 2010.

[3] S. Sridhar, "Extended Angel: Knowledge-based Approach for LOC and effort Estimation for Multimedia Projects in Medical Domain", International Journal of Software Engineering & Applications (IJSEA), Vol.2, No.4, October 2011.

[4] P. A. Whigham, C. A. Owen, andS. G. Macdonell, "A baseline model for software effort estimation". Journal ACM Transactions on Software Engineering and Methodology , ACM TOSEM,Volume 24 Issue 3, May 2015 PP 20:1–20:11, 2015.

[5] L. M. Alves, S. Oliveira , P. Ribeiro,R. J. Machado, "An empirical study on the estimation of size and complexity of software applications with function point analysis", 14th International Conference on Computational Science and its Applications; Guimaraes, IEEE 2014. p. 27–34.

[6] M. Harizi, "The Role of Class Diagram in Estimating Software Size", International Journal of Computer Applications, Volume 44, No. 5, pp. 31-33, 2012.

[7] K. U Sharani, V. Vignaraj Ananth., D. Velmurugan, "A survey on software effort estimation", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, India, 2016.

[8] J. Zivadinovic, Z. Medic, D. Maksimovic, A. Damnjanovic, S. Vujcic, "Methods of Effort Estimation in Software Engineering", International Symposium Engineering Management and Competitiveness 2011 (EMC2011) June 24- 25, 2011, Zrenjanin, Serbia.

[9] M. Azzeh, and A. B. Nassif, "Fuzzy Model Tree For Early Effort Estimation", 12th International Conference on Machine Learning and Applications, DOI 10.1109/ICMLA.2013.115, IEEE computer society, 2013.

[10]A. K. Bardsiri, and S.M. Hashemi, "Software Effort Estimation: A Survey of Well-known Approaches", International Journal of Computer Science Engineering (IJCSE), 2014. 3(1): pp. 46-50.

[11]M. Saroha, S. Saha; "Analysis of various Software Effort Estimation Techniques", International Research Journal of Computers and Electronics Engineering (IRJCEE) Vol. 3, Iss. 2 , 2015.

[12]M. M. Kirmani, and A. Wahid," Revised Use Case Point (Re-UCP) Model for Software Effort Estimation", International Journal of Advanced Computer Science and Applications, 6, 65-71,2015.

[13]M. M. Kirmani, and A. Wahid, "Impact of Modifications made in Re-UCP on software effort estimation" Journal of software engineering and applications, Vol. 8, No. 6, pp276-289,2015.

[14] P. K. Bhatia, G. Kumar, " Role of Technical Complexity Factors in Test Effort Estimation Using Use Case Points ", International Journal of Software Engineering Research & Practices Vol.1, Issue 3, July, 2011.

[15] R. K. Clemmons, "Project Estimation With Use Case Points", The Journal of Defense Software Engineering, February 2006.

[16] http://sparxsystems.com.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Design and implementation elliptic curve digital signature algorithm using multi agent system

Tawfeeq M. Tawfeeq Al-Flaih
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
Tawfeek.mflaih@yahoo.com

Marwa Adeeb Al-jawaherry
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
marwaaljoharjy123@gmail.com

Abstract— **The rapid and growing development of information and communication technologies ICTs, especially in the Internet, has been a key driver for improving the quality and efficiency of services provided by many countries. The digital signature algorithm (DSA) is designed to dispense with the signature in handwriting and replace it with a signature, and it helps us to verify the identity of the sender and receiver in a reliable and secure manner.**

In this research we are proposing and constructing security system which is called Digital Signature Multi Agents (DSMA). It is based on Multi Agent System (MAS) and provides authentication of senders or receivers by applying "Elliptic Curve Digital Signature Algorithm (ECDSA)" to sign and to verify the electronic documents. Two types of agents were developed in our proposed system: sender and receiver agent. Java programming language and JADE (Java Agent Development framework) were used to constructing DSMA.

Keywords- Hashing Algorithm; Elliptic Curve Cryptosystems; Elliptic Curve Digital Signature Algorithm; Multi Agent System; Java Agent Development

I. INTRODUCTION

With the increasing of the online application and electronic transactions, The transition from paper based transactions to electronic transaction become more easy and less complicated but the challenge lies in the implementation of these transactions in terms of the validation and insurance. It can be viewed to the digital signature technology as a mechanism to maintain the integrity and safety ratio in electronic transaction [10]. As the dependence on the Internet for the exchange of information and communication continues to increasing, the security concerns are becoming more important. There is desperately need for digital identity or digital signature, which will activate the quality of our dealings and contacts increase the security [17].

It is noted that when conducting transactions electronically, there is no way to confirm the identity of the sent or received transections, hence the possibility of the use of digital signatures to authenticate the source of the electronic messages or transactions; The digital signature confirms the true identity of the sender, and more importantly, can be used to maintain the integrity data, from editing, sources posing a strength and excellence, for these reason digital signatures is an effective solution for authentication and documentation [14][23].

In this paper we built and developed a secure system based MAS called Digital Signature Multi Agents (DSMA) that sends text messages to many distributed sites, and implement a Digital signature algorithm (DSA) to verify the integrity of sent and signed data, in addition to verifying sender identity.

DSMA can be executed on any system that rely style electronic exchange of official documents, It is possible to apply DSMA system for the exchange of official document traded electronically between the presidency of the University of Mosul and between different colleges or for the exchange of official document electronically between Deanship colleges and departments. DSMA system can accept any number of users, the user can be divided into two types:

A. Sender

The person who is sends electronic documents after the process of generating a digital signature and encryption of the document and then sends it to the receiver site by his personal agent.

B. Receiver

The person who is receives electronic documents that have been digitally signed and transmitted by sender agent.

It is worth mentioning that all users of DSMA system can be mailed electronic documents at the same time.

II. MOTIVATION

The digital signature (DS) is a mathematical method to clarify whether the digital messages or documents received are true or not. It also helps the receiver to verify the validity of the sender's identity (authentication), in this case, the sender cannot deny sending the message (non-repudiation), and can be sure that the message is not changed during the transfer process (integrity).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

DS is one of the standard elements in most suites of the cryptographic protocol, DS is used in many areas, such as financial dealing, software used in distributions and contract management systems, As well as systems to detect tampering or counterfeiting. Multi-agent systems are used in many areas such as network security systems [5].

Our research aims to develop security system based on MAS, We tried to overcome some of the difficulties we faced in this paper, recognize the authentication of the signatory and using multi agent system to implement. Our system has several requirements that have to be achieved in his work as follows:

1. The main function of the system is to make sure of the identity of the sender of confidential information after receiving it.
2. The number of system user is different and is not specific to a certain number, for example there may be four or five users.... etc., and so the number of computers linked to the network is not specified.
3. Any user of the system can send information over the network to the rest of other users through the software agent.
4. Every user of the system has a personal agent represents him and who interacts with each others by sending messages.
5. Any user can be the sender and receiver of the messages in the future and at the same time.

III. RELATED WORK

As long as people have been able to communicate with one another, there has been a desire to do so secretly. Many researchers work with digital signature algorithm, Cloud Computing and MAS:

1. In (2011), Erfaneh Noorouzi1 and his colleagues proposed a new DSA algorithm, which generates dynamic size hash files, which mean the size of the message affects the result of the hash function. The mechanism for (hash / encrypt) will be more simple by a new DSA algorithm [18].

2. In (2011), Aarti Singh and his colleagues proposed "security engine" to secure messages sent in networks environment, this proposed make Elliptical Curve keys used for the purpose encrypt and decrypt. This framework can be implemented in the security layer of the current wireless communication model for this reason is not needed to rewrite it to use [29].

3. In (2012), Salwanibtmohd Mohd daud and his colleagues produced DS from achieved a simple mechanism by proposing a new algorithm. The resulting output would be dynamic and smaller by this new algorithm. Hashing and encoding the message after the algorithm read the input file [19].

4. In (2012), Thulasimani Lakshmanan and Madheswaran Muthusamy used SH Algorithm to present a new SHA called "SHA-192". The output length message "SHA-192" of 192. They designed "SHA-192" to resist the SH Algorithm attacks and to fulfill the different level of information security[14].

5. In (2016), Virangna Pal and his colleagues discussed the two types of Security algorithms (Symmetric and Asymmetric Algorithms) that were used in "Cloud Computing", they checkup various constraints for ex: features and mechanisms, and they discussed some case connected with distributed system [30].

IV. METHODS & MATERIALS

A. *Electronic Signature*

An electronic signature refers to data that has an electronic format, which is associated to other electronic data logically, and this data will be used by the signatory to perform signing process. The main objective of applying the electronic signature process is to provide accurate and safe way to verify the identity of the sender. It is worth noting that the definition of electronic signature depends on the jurisdiction that Applied. There are three types of electronic signature, as following: Digital Signature, Personal Signature and Signature Using Pen Mail [4] [16].

Digital signature is an encryption process is composed of some of the letters, symbols and numbers. It can be represented as a string of binary digits in a computer, and must achieve the functions where the signature identifies the signer's identity and the expression of his will approve the content of the message data [23][12]. The digital signature value is calculated using a number of parameters that verify the integrity of the signed data and the identity of the signatory [10] [13]. The digital signature having several requirements includes UN forgeable, User authentication, Non-repudiation, Unalterable and Not reusable [20].

B. *Software Agents*

Agents are separate pieces of software that have the ability to act independently and interact with the environment in which they operate. There are different types of agents so their abilities are also different. In order for the agent to be described as an "intelligent" agent, he should have the ability to interact with other agents or with his environment without the need for direct interaction by human beings as well as must be flexible [6].There are four types of agents: Executive agents, Collaborative agents and Contributory agents [9][25].

Multi-agency systems are modern approaches to analysis, design and implementation of complex software systems. To develop and implement different types of software systems, it is possible to use multi-agent systems and is also used in the development of search and rescue systems and network security [11]. MAS are used to describe several agents that interact with each other positively, but also negatively within an environment [22][8].

C. *JADE platform*

JADE platform is a software framework port language Java. It was developed by the Research Institute of the Italian contact in 1998 by using a set of graphical tools to simplify the implementation of multi-agent systems [26]. The goal of JADE is to facilitate the development and to ensure that the standard

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

response by providing a set of services for the overall system, as well as providing a variety of agents [27].

JADE architecture consists of agent containers that are on the same platform but distributed over the network. Each agent lives in a container which is a Java process that provides a JADE runtime and all services necessary to host and execute agents. In each platform there is a special container, called the main container, which is launched at the platform and which contains the other containers in which they are registered [22]. The interaction is the most important properties of the agent, and the agent interaction to share information and knowledge in order to achieve his goals. In order every agent to own a mechanism to achieve compatibility, there are two key elements in the agent connections: Protocol negotiations common/ language of communication and Representation of the general formula content [27].

D. Secure Hash Algorithm (SHA)

Secure Hash Algorithms provides many services while used in other cryptographic algorithms [1]. Converting a variable length message into a condensed representation of the electronic data in the message is made by Hash algorithm, and this output can then be used for DS and any other secure system. When employed this representation in a DS application, the "Hash value" of the message is signed instead of the message itself, then the receiver can authenticate the integrity of the signed by using the signature to verify the signer of the message [3][18][21].

E. Digital Signature Algorithm (DSA)

The U.S.NIST in August 1991 proposed "Digital Signature Standard (DSS)" [23]. The key generation process consists of two stages, choosing algorithm parameters that can be shared between the various users of the system, in the first stage. The second stage involves calculating the private and public keys for the same user [17].

F. Elliptic Curves Cryptosystems (ECC)

In 1985 Neal Koblitz and Victor Miller invented ECC. It can appear as EC analogues of the older Discrete Logarithm crypto systems [15]. Public key cryptography based on the algebraic structure of EC over finite fields. ECC requires smaller keys compared to any other cryptography to provide equivalent security. ECC are applicable for (key agreement, DS, and other tasks), they can be used for encryption by combining the key agreement with a symmetric encryption scheme and used in several integer factorization algorithms based on ECC [28].

G. Elliptic Curve Digital Signature Algorithm (ECDSA)

The ECDS Algorithm is the "Elliptic Curve" analogue of the commonly used DSA [20]. ECDS Algorithm offers technical avail in the areas of certificate, performance, and key over other DS methods [24], Figure (1) shows the interaction between ECDS Algorithm and SHA-2 [7]. RSA or DSS are very difficult or expensive to implement in specific applications while smaller data structures and calculation efficiencies for ECDS Algorithm enable it to be used in these applications [13][16].

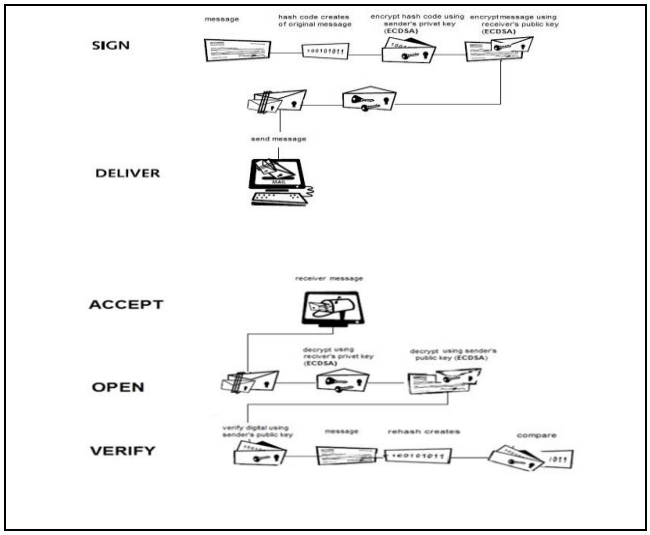


Figure 1. Represent interaction between ECDSA and SHA-2

V. THE LIFE CYCLE OF PROPOSED WORK

This paragraph presents clarification of the proposed DSMA system, which is designed to ensure and confirm the identity of the sender of electronic documents, as well as making sure it from the correct source. We will describe the system architecture, as well as explain JADE interfaces that are used to communicate with the system and clarify all code components. Also we will indicate the number of agents in DSMA system and user characteristics and responsibilities. We will use the Smart MAS style to analyze, the design and implementation of our system.

A. Requirements phase

After the initial analysis of the requirements, the representation of the active ingredients in the simple scheme actor. In our proposed system we have two types of actors, first: sender of the messages, second: receiver of the messages. The main objective of the sender is to generate digital signature and send messages. While the main objective of the receiver is receiving the messages and confirming the identity of the sender, Figure (2) shows the actors of a simple scheme for DSMA system.

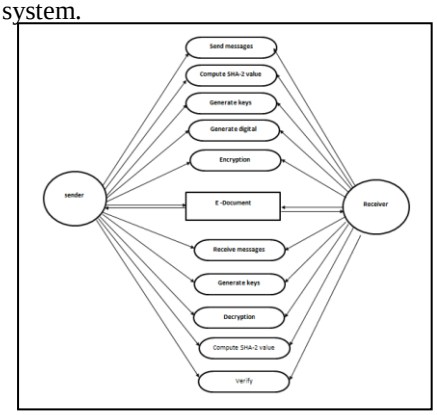


Figure 2. Simple scheme actor of the DSMA system

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Advanced requirements consist of four steps: insert system actors, creating goals diagrams, creating actor diagrams and analyze dependencies.

1. Insert system actor: in this step the system actors are inserted under development in a simple diagram and its own tasks are appointed as shown in Figure (3), which shows the system actors that have been delegated all the goals except the resources which are outside the system.

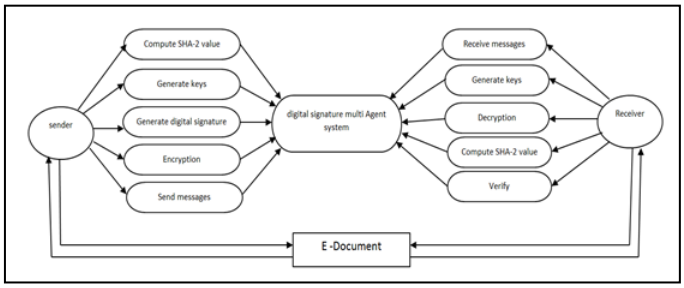


Figure 3. illustrates the DSMA system Actor

2. Creating goals diagrams: This step is centered in three sub-phases, As explained In Figure (4) and Figure (5). show the goals decomposition of the sender and receiver.

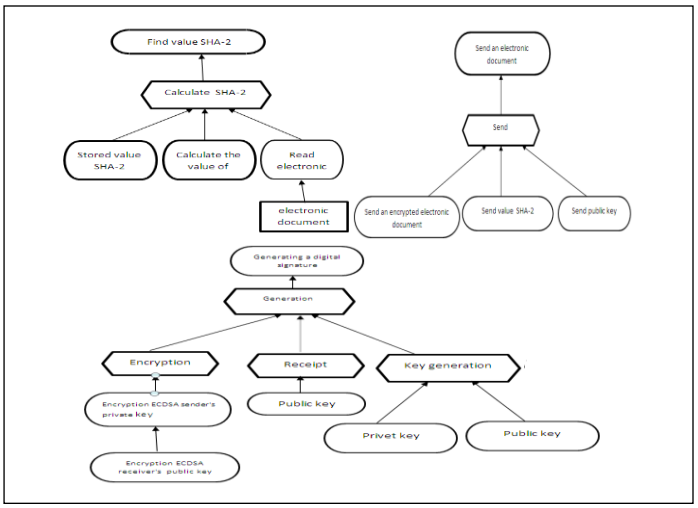


Figure 4. Goals decomposition of the sender

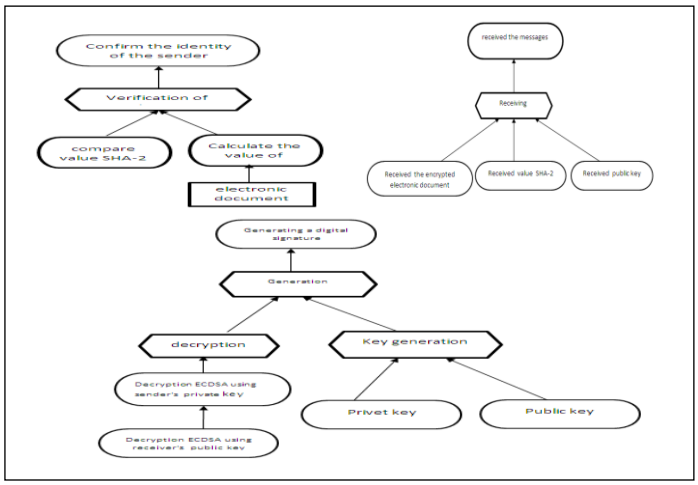


Figure 5. Goals decomposition of receiver

3. Creating actor diagram: After assembling the plans for system actors and goals, final actor diagram is formed for requirements phase, Figure (6) shows the final actor diagram.

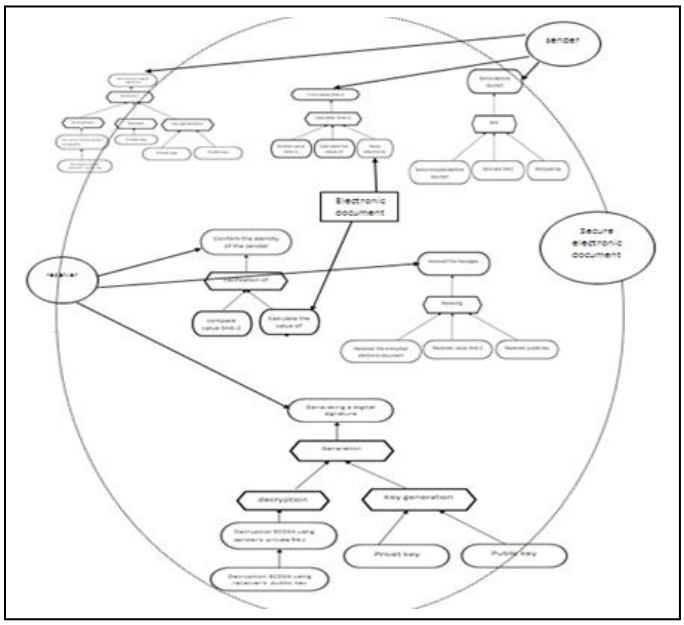


Figure 6. Final actor diagram

4. Analyzed dependencies: it was analyzed between actors who are (sender, receiver) and DSMA system as shown in table (1).

B. *Analysis Phase*

The analysis phase is divided into two major steps: the first is to create a structure description of the dependencies between agents, see table (1). And the other step is a description of the role for each agent in DSMA system, see table (2).

TABLE I. ANALYZED DEPENDENCIES BETWEEN ACTORS IN DSMA

Sender dependencies It is needed sender actor to achieve its goals, as in the following formulas: **Dependency**: find the value of hash **Dependent**: sender. **Dependee**: receiver. **Dependum**: Electronic document. **Goal**: calculate the hash value by implementing the algorithm sha-2 (384-bit). **Pre-condition**: the presence of the electronic document. **Post-condition**: determining the validity of information and use it to generate DS. **Dependency**: the generation of the DS. **Dependent**: sender. **Dependee**: receiver. **Dependum**: Electronic document. **Goal**: generates a digital signature for electronic document. **Pre-condition**: the presence of the electronic document and find the value computed hash. **Post-condition**: generating each of the (public, private) keys and receiver of receiver public key, and encrypt electronic document **Dependency**: electronic document using ECDSA. **Dependent**: sender. **Dependee**: receiver. **Dependum**: Electronic document. **Goal**: encrypt electronic document. **Pre-condition**: the existence of a signed electronic document. **Post-condition**: Send electronic document to the receiver. **Dependency**: Sends an electronic document. **Dependent**: sender. **Dependee by**: receiver. **Dependum**: Electronic document. **Goal**: electronic document delivery to the receiver. **Pre-condition**: the existence of an encrypted electronic document and provide contact with the receiver. **Post-condition**: Posted / / Not transmitter.	**Receiver dependencies**: It is needed receiver actor to achieve its goals, as in the following formulas: **Dependency** : receive of messages. **Dependent**: receiver. **Dependee**: sender. **Dependum**:messages. **Goal**: receive of the electronic document and the public key and the value of the SHA-2. **Pre-condition**: the presence of the electronic document. **Post-condition**: has been receiving / / No receive. **Dependency**: generate keys **Dependent**: receiver. **Dependee**: sender. **Dependum**: generate the (public, private) key . **Goal**: electronic document will be encrypted using the public key by the sender and use the private key to decrypt the electronic document by the receiver. **Pre-condition**: the presence of the receiver. **Post-condition**: Send the public key to the sender. **Dependency**: decryption of electronic document . **Dependent**: receiver. **Dependee**: sender. **Dependum**: message. **Goal**: decrypt message using ECDSA. **Pre-condition**: receive the electronic document from the sender **Post-condition** : perform verification algorithm. **Dependency** : calculate the value sha-2 (384-bit). **Dependent**: sender. **Dependee**: the receiver. **Dependum**: Electronic document. **Goal**: find the value of SHA-2. **Pre-condition**: electronic document encryption using ECDSA. **Post-condition**: compare the value of sha-2 calculated with the value of SHA-2 received from the sender. **Dependency** : confirm the identity of the sender of the electronic document . **Dependent**: receiver. **Dependee**: sender. **Dependum**: electronic document. **Goal**: confirm the identity of the sender by verifying the digital signature . **Pre-condition**: compare the value of sha-2 with the calculated value of the SHA-2 received from the sender. **Post-condition**: Sender is trusted / / sender is not trusted.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II. CLARIFIES THE ROLE OF THE SENDER AND RECEIVER

The role of the sender	The role of the receiver
Description: This role is a process of sending electronic document after calculating the SHA -2 , encrypted it using ECDSA algorithm and confirms the identity of the sender when it is sents . **Main Goal**: generates a digital signature and send electronic document . **Dependency** : Send an electronic document . **Activities**: receives the receiver's public key and send electronic document , application two algorithms SHA-2 , and ECDSA and configure agents. **Successful actions**: the generation of the digital signature . **Failed actions**: the inability to generate a digital signature, send electronic document , and verify the identity of the sender.	**Description**: This role is a process of receiving electronic document which is sent by the sender, decryption using ECDSA algorithm, and calculate the value of the SHA-2 to make sure of the identity of the sender of the electronic document . **Main Goal** : receives the electronic document , and confirm the identity of the sender. **Dependency** : receive the electronic document , the sender's public key , the value of the SHA-2. **Activities** : receives electronic document , send the public key of the receiver to the sender ,applying SHA-2, and ECDSA algorithms, comparison and configure agents. **Successful actions:** the exchange of messages and communicate with the sender. **Failed actions**: the inability receive electronic document . and can't verify the

There are two types of agents, each one with a specific role in the DSMA system Figure (7) illustrates agents and its own role.

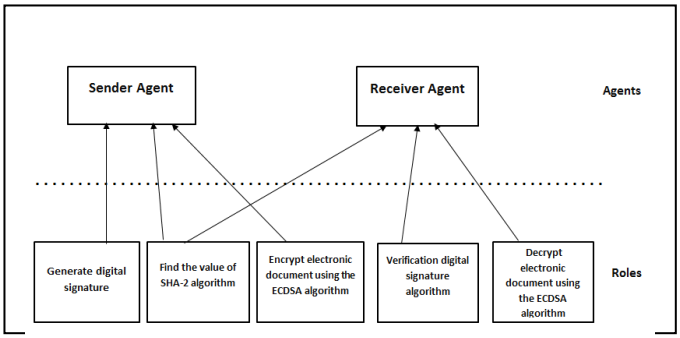


Figure 7. Agents and its own role

C. *Design Phase*

After the defining of agents and setting goals and their own tasks, we can create a plan to deploy these agents in locations that can be found there in, as well as a description of its functions. The proposed system contains two of the agents who are the sender and the receiver and the note through the requirements, it should be distributed and can be There a different numbers of senders and receivers, On this basis deployment scheme has been configured the presence of the sender in the platform and the agent of the receiver in another platform exist on the same network and the number of copies of the sender agent and receiver agent is not specified because it is dependent on the number of users of the system. As a software engineer we are focus on implementing the most important design concepts that allied "Modularity" which divided software to components, each component has its own name and address that called "Modules". Figure (8) show the control hierarchy and the modules in DSMA system.

D. *Construction Phase*

DSMA system consists of one package that contains the following classes: Sender Agent Class, Receiver Agent Class, Encryption Class, Decryption Class, Digital Signature Class, Hashing Class. We built the DSMA system using the JADE framework under Java language.

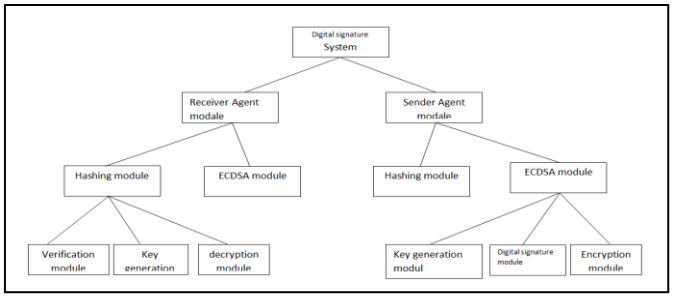


Figure 8. control hierarchy and modules in DSMA system

VI. CASE STUDY

This section presents the using and testing of DSMA system. It implemented practically and discussed according to the results obtained.

The proposed DSMA system is a distributed system, so the implementation needs to provide a number of computers linked with each other through the LAN, and the number of those computers is not specified, and each user can interact with other users through the agent who represented him and the agent will keep working in his computer.

In the beginning, DSMA system is used by running the JADE platform at the sender and receiver sides, then the user should create sender and receiver agent on each platform. The sender agent reads the electronic document and finds the value of the hash code using algorithm SHA-2 (384-bit) and sends it to receiver, the keys and digital signature are to be generated. Then encrypts electronic message and sends it also to receiver. The receiver agent receives electronic messages and decrypts it, finds the value of hash code using the algorithm SHA-2, compares the value of hash calculated with the value of the hash received, if the value is equal, it means that the document is received from the correct sender, but if it is not equal, it means that the document is received from the incorrect sender. On this basis, assurance of the identity of the sender is achieved. Figure (9) and figure (10) show the JADE interface at Sender / Receiver site.

Agents are interacting with each other by exchanging ACL messages, several types of behaviors are used in the implementation of the tasks of the agent, namely: One shot behaviors and cyclic behaviors. Sender and receiver agents can be resident in the main or secondary containers in JADE platform.

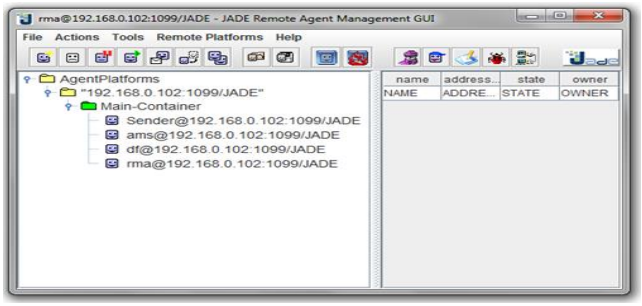


Figure 9. Sender agent.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

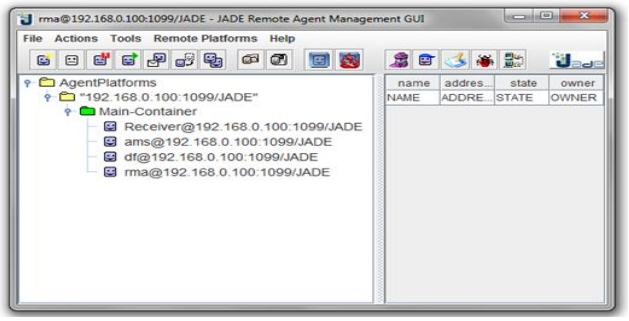


Figure 10. Receiver agent

DSMA system includes two types of agents:

A. The sender's agent:

1. (Key Generation Algorithm):

At this step the public key and private key are generated. Later keys are to be used in encryption and decryption operations.

```
.
.
public class Key_Generate {
        public static String[] point(String s1, String s2 , String l)
        {  BigInteger a,b,d,x1,y1,s,aa,bb,v1,v2,c,c1,x,y;
                                .
                                .
                                .
public static String[] key_generater_pri(String private_key) throws
UnsupportedEncodingException
        {
          BigInteger q1,q2;
          String[] result_A;
             result_A = point("0","2",private_key);
             q1= new BigInteger(result_A[0]);q2= new
BigInteger(result_A[1]);
              String ar[] = new String[2];
             ar[0]=q1.toString();
         ar[1]=q2.toString();
           return ar;
         }
}
```

2. (Signing Algorithm):

By using this algorithm the sender agent generate a digital signature for electronic document and implicitly calculates the value of the SHA-2.

3. (Key exchange operation):

The sender agent sends his public key to receiver agent.

```
public class Sender extends Agent
{ String str,hash,result[],result1[],result2[];
String result_pub_A[],result_pub_B[],priv_A,priv_B;
protected void setup()
                        {
                                System.out.println("Enter  String  :
");
                                Scanner       sc11       =       new
Scanner(System.in);
                            str = sc11.nextLine();
                        System.out.println("Enter Private Key A : ");
                        priv_A=sc11.nextLine();
                                try {
 result_pub_A=Key_Generate.key_generater_pri(priv_A);
                                }
                        System.out.println("public
key"+result_pub_A);
 addBehaviour(new CyclicBehaviour(this)
                        {               public void action() {
                                        ACLMessage msg1rec= receive();

                                                        if (msg1rec!=null){
                                                                String title =
msg1rec.getContent();


        System.out.println( " - " +

myAgent.getLocalName() + " <- " +

msg1rec.getContent() +"    " + title);


                                                        block();
                                                        }
                                }
                };
```

4. (Encryption Algorithm):

At this step electronic document was encrypted by applying ECDSA algorithm before sending them to the site of the receiver, the encryption of electronic document contain two operations: First, encrypted electronic document by using private key of the sender , second, encrypt the electronic document by using the public key of the receiver.

```
public class encryption {
        public static String[] point(String s1, String s2 , String l)
        {
                        BigInteger a,b,d,x1,y1,s,aa,bb,v1,v2,c,c1,x,y;
                        BigInteger p = new BigInteger("137");
                        aa= new BigInteger("3");
          bb= new BigInteger("2");
          c= new BigInteger("4");
          c1= new BigInteger("27");
          a = new BigInteger("1");
          b = new BigInteger("4");
          x = new BigInteger(s1);
          y = new BigInteger(s2);
          d = new BigInteger(l);
          int ss = p.intValue();
          ss=ss-2;
                        x1=x;
          y1=y;
                        for(int i=2;i<=d.intValue();i++)
          {
                                if (x1==x&&y1==y)
            {

        v1=x.pow(2).multiply(aa).add(a);
                                v2=bb.multiply(y);
                                v1=v1.mod(p);

s=v2.pow(ss).multiply(v1);
                                                s=s.mod(p);

x1=s.pow(2).subtract(x1).subtract(x);
                                x1=x1.mod(p);

y1=s.multiply(x.subtract(x1)).subtract(y);
                                y1=y1.mod(p);
                }
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
                                        .
                                        .
                                        .
public static String[] encrypt(String str1 ,String x,String y,String pri)
throws UnsupportedEncodingException
        {
         String[] result;String[] result_key,decrypt,re = null;
        String cipher2 = null;
         String stre = null,strd = null,ad;ad=" ";
         int[]a;int c=0;int e=0;
         int l;
         l=str1.length();
         if(l%2==0)
         {
                        System.out.print(" ");
         }
         else
         {
                        str1=str1+ad;
         }
         BigInteger q1,q2,c1,c2,s1,s2,d1,d2,v,p,c21,c22,kk;
         v = new BigInteger("-1");
         p = new BigInteger("137");
         int co=0;int er=0;
         byte[] bytes = str1.getBytes("US-ASCII");
         l=str1.length();
         result_key=point(x,y,pri);
         s1=        new        BigInteger(result_key[0]);s2=        new
BigInteger(result_key[1]);
         BigInteger r,r1;String strre = null;
         String re1 = null;
         String[] re2;
         while(c<l)
         {              r = new BigInteger(Byte.toString(bytes[c]));
                        r1 = new BigInteger(Byte.toString(bytes[c+1]));
        cipher2=add2point(r.toString(),r1.toString(),s1.toString(),s2.to
String());
                        strre=strre+cipher2;
                        c=c+2;
         }
         strre=(String) strre.subSequence(4, strre.length());
         strre=(String) strre.subSequence(0, strre.length()-1);
         String[] bytes1 = strre.split(",");
         System.out.print("Encryption:");
         for (int a1=0;a1<bytes1.length;a1++)
         {        stre=stre+bytes1[a1];

System.out.print((char)Integer.parseInt(bytes1[a1]));
         }
          String ar[] = new String[4];
           ar[0]=stre;
           ar[1]=strre;
                        return ar;
         }}
```

TABLE III. ILLUSTRATE THE TYPE OF AGENT, MESSAGES INFORMATION (TYPE, NUMBER, AND CONTENT) IN SENDER AGENT.

Name of agent	Type of agent	No. of sending messages	No. of receiving messages
Sender	Static	3	1
In	**Messages Content**		**Message type**
1.	Public Key		Send
2.	Value of SHA-2		Send
3.	Signed electronic document		Send
4.	Receiver Public Key		Receive

B. The receiver's agent:

The receiver agent is responsible for receiving and decrypting electronic document and confirms the identity of the sender using ECDSA algorithm which is implicitly consists of three algorithms that are implemented at the receiver site as follow sequence, as can be seen in table (4).

1. (Key Generation Algorithm):

Public key and private key are generated. Keys will be used in encryption and decryption operations.

2. (Key exchange operation):

The receiver agent sends his public key to sender agent.

3. (Decryption Algorithms):

The decryption of electronic document contains two operations: First, decrypting electronic document by using private key of the receiver, second, decrypting the electronic document by using the public key of the sender.

4. (Signature verification algorithm):

Through using this algorithm we can ensure the authenticity of the digital signature after decrypted, and find the hash value of electronic message, then compare it with the hash value received from the sender's agent to ensure the identity of sender.

```
public class Receiver extends Agent
{
       String str,hash,result[],result1[],result2[];
       String result_pub_A[],result_pub_B[],priv_A,priv_B;
       private static final long serialVersionUID = 1L;

       protected void setup()
   {
                      Scanner sc22 = new Scanner(System.in);
                      System.out.println("Enter Private Key B : ");
    priv_B=sc22.nextLine();
    try {
           result_pub_B=Key_Generate.key_generater_pri(priv_B);
        }
    System.out.println(result_pub_B);

    addBehaviour(new OneShotBehaviour(this)
                      {
                       public void action() {
                       ACLMessage        msg2rec        =        new
ACLMessage(ACLMessage.INFORM);
                              final     String                    s1=
result_pub_B[0]+"**"+result_pub_B[1];
                                msg2rec.setContent(s1);


msg2rec.addReceiver( new AID( "s" , AID.ISLOCALNAME) );
                                send(msg2rec);
                                }
                      });
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE IV. ILLUSTRATE THE TYPE OF AGENT, MESSAGES INFORMATION (TYPE, NUMBER, AND CONTENT) IN RECEIVER AGENT.

Name of agent	Type of agent	No. of sending messages	No. of receiving messages
Receive	Static	1	3
In	**Messages Content**		**Message type**
1.	Public Key		Send
2.	Value of SHA-2		Receive
3.	Signed electronic document		Receive
4.	Receiver Public Key		Receive

VII. CONCLUSION

The present study, proposed a secure multi agent system (DSMA), and many points have been concluded besides the following:

1. The software agent has an ability to execute complex algorithms, and excellent results were got .

2. The ssoftware agent was a very good choice to execute numerical algorithms efficientlyy.

3. MAS has an ability to reduce the communication problems because the of low size of agent's messages.

4. The use multi agent system help to perform complex interaction between distributed sites.

VIII. ACKNOWLEDGMENT

We thank the in charge of laboratories of the Software Engineering Department - Mosul University for their help in testing DSMA system propose.

IX. REFERENCES

[1] Zhu, H., and Li, D.,(2008), "Research on Digital Signature in Electronic Commerce", Internationa MultiConference of Engineers and Computer Scientists, 19-21 March,Vol.1, IMECS 2008, Hong Kong.

[2] Bellifemine, F., Poggi, A., and Rimassa, G.,(2001), "Developing multi – agent systems with a FIPA-compliant agent framework", Softw:Pract.Exper., 31:103 –128.doi:10.1002/1097-024X(200102)31:2<103::AID-SPE358>3.0.CO;2 O.

[3] Bellovin, S., and Rescorla, E.,(2005), "Deploying a New Hash Algorithm", Technical Report, National Institute of Standards and Technology (NIST).

[4] Bettelli A.V.,(2004), "The Trustworthiness in Software Agents' Electronic Signatures".

[5] Borselius, N.,(2003), "Multi-agent system security for mobile communication", PhD Thesis submitted to the Department of Mathematics Royal Holloway, University of London.

[6] Borselius, N,.(2002), "Mobile Agent Security", Electronics & Communication Engineering Journal, Volume: 14, Issue: 5.

[7] Curry, I.,(2001), "An Introduction to Cryptography and Digital Signatures", Version 2.0, Entrust Securing Digital Identities & Information.

[8] Dale, J.,(1997), "A Mobile Agent Architecture for Distributed Information Management", PhD thesis, University of Southampton.

[9] Fabio, B., Caire, G., and Greenwood, D.,(2007), "Developing Multi-Agent Systems with JADE", John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, ISBN: 978-0-470-05747-6.

[10] Gallagher P. and Director D.,(2009), "Digital Signature Standard", National Institute of Standards and Technology, Gaithersburg, MD 20899-8900 , FIPS PUB 186-3.

[11] Hegde, M., and Singh S.,(2013), "Alert-BDI: BDI Model with Adaptive Alertness through Situational Awareness", IEEE. Conference Mysore, India. ISBN: 978-1-4673-6217-7

[12] Johnson D., Menezes, A., and Vanstone, S.,(2001), "The Elliptic Curve Digital Signature Algorithm (ECDSA)", International Journal of Information Security, Volume 1, Issue 1, pp 36–63

[13] Khalique, A., Singh, K., and Sood, S.,(2010), "Implementation of Elliptic Curve Digital Signature Algorithm", International Journal of Computer Applications, (0975 – 8887), Volume 2 – No.2.

[14] Lakshmanan, T., and Muthusamy, M.,(2009), "A Novel Secure Hash Algorithm for Public Key Digital Signature Schemes", The International Arab Journal of Information Technology, Vol. 9, No. 3, May 2012.

[15] Leslie, M.,(2006), "Elliptic Curve Cryptography", 40575063, Advanced Combinatorics.

[16] Liao, H., and Shen, Y.,(2006), "On the Elliptic Curve Digital Signature Algorithm", Tunghai Science, Vol. 8: 109−126.

[17] Muthukuru, J., and Sathyanarayana, B.,(2012), "A Secure Elliptic Curve Digital Signature Approach without Inversion", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-3, Issue-2.

[18] Noorouzi1, E., HAGHIGHI, A., Peyravi, F., and zadeh, A.,(2011), "A New Digital Signature Algorithm", 2009 International Conference on Machine Learning and Computing, IPCSIT vol.3 (2011), IACSIT Press, Singapore.

[19] Noroozi, E., DAUD, S., SABOUHI, A., and NAMADI, M.,(2012), "New Implementation of Hashing and Encoding in Digital Signature", 2012 IACSIT Hong Kong Conferences, IPCSIT vol. 29 IACSIT Press, Singapore.

[20] Rached, T., Mohsen, M., Noura, B., and Wajih, E.,(2012), "Low Power Elliptic Curve Digital Signature Design for Constrained Devices", International Journal of Security (IJS), Volume (6) : Issue (2).(1).

[21] Rjasko, M.,(2008), "Properties of Cryptographic Hash Functions", Cryptology Archive, ePrint, Report 2008/527.

[22] Shehory, O.,(1998), "Architectural Properties of Multi-Agent Systems", The Robotics Institute, Carnegie Mellon University - Pittsburgh, Pennsylvania 15213.

[23] Tonien, D., To, V., and Seberry E.,(2011), "A New Generic Digital Signature Algorithm", Groups, Complexity, Cryptology, 3 (2), 221-237.

[24] Brown, D.,(2009), "SEC 1: Elliptic Curve Cryptography", STANDARDS FOR EFFICIENT CRYPTOGRAPHY, Version 2.0.

[25] Wang, A.,(2001), "Using a Mobile, Agent-based Environment to support Cooperative Software Processes", M.Sc., Norwegian University for Science and Technology, Department of Computer and Information Science.

[26] Yang, Q.,(2008), "A Multi -agent Prototype System for Helping Medical Diagnosis", Thesis (Masters), Memorial University of Newfoundland.

[27] Zhang, C., Xi1, J., and Yang, X.,(2008), "An Architecture for Intelligent Collaborative Systems based on Multi-Agent", IEEE, 978-1-4244-1651, Official U.S. Government information about the Global Positioning System (GPS) and related topics.

[28] U.S. National Security Agency.,(2016), "Information Assurance Directorate", Commercial National Security Algorithm Suite and Quantum Computing FAQg.

[29] Singh, A., Juneja, D., and Sharma, A.,(2011), "Elliptical Curve Cryptography Based Security Engine for Multiagent Systems Operating in Semantic Cyberspace", International Journal of Research and Reviews in Computer Science (IJRRCS),Vol. 2, No. 2.

[30] Pal, V., Chowdary, P., and Gaur, P.,(2016), "A Review of Security Algorithms in Cloud Computing", International Journal Of Science & Technoledge (ISSN 2321 – 919X), Vol. 4, Issue. 10.

Firefly Algorithm Implementation Based on Arduino Microcontroller

Riyadh Zaghlool Mahmood
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
riyadh_1970@yahoo.com

Marwa Adeeb Al-jawaherry
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
marwaaljoharjy123@gmail.com

Abstract— There are many optimization algorithms; "Bio inspired metaheuristic algorithms" are one of the most important optimization algorithms. In this research, we intend to implement "Firefly Algorithm (FA)", which is one of the "Bio inspired metaheuristic algorithms" to optimize the finding operation of the maximum and minimum values of various mathematical equations based on Arduino microcontroller. The results are displayed on the GLCD, the following information is displayed: the number of the iteration (Itre), the minimum value (x), the maximum value (y) of variables in mathematical equivalents, the value of lightness (l), and finally the value of error (E).

Keywords: Optimization, Firefly Algorithm, minimum and maximum values, mathematical equations, Arduino mega2560.

I. INTRODUCTION

In most engineering and scientific problems, optimization is one of the most important ways to solve it, and through the continuous development in recent years many methods of optimization developed to optimize the solving of these problems. The most public methods are the metaheuristics methods [1].

At present one of the most common algorithms in global optimization problems is the nature-inspired metaheuristic algorithms", especially NP hard optimization. An example of those algorithms is the Swarm Optimization algorithm, developed in 1995 by Kennedy and Eberhart; these algorithms relied on the behavior of natural systems such as the bird schooling and fish. This algorithm was recently applied to find optimal solutions for many optimization applications [19].

The first source of inspiration for the design and development of many new optimization problems is the behavior of natural systems, such as ants systems, which is developed by observing the nature of ants system in nature, swarm intelligence is the behavior applied by these algorithms. It is therefore dependent on the interaction of individual entities and its social behavior is inspired from the behavior followed by insects [12].

The firefly algorithm was developed by the "Xin-She Yang", a firefly algorithm inspired by the behavior of fireflies in nature, two thousand firefly species is the estimated number of the their population. Most of these fireflies produce rhythmic and short flashes. Bioluminescence process generates flashing light of the fireflies. It may serve as warning signals or an element of court ship rituals [18].

In this research we choose to design and implement the firefly algorithm to find maximum and minimum values of mathematical equations, Arduino microcontroller was used to develop our proposed system and the system results were displayed on GLCD.

II. RELATED WORK

Bidar M. and Kanan H. R. [4] proposed an algorithm inspired from Firefly algorithm. The researchers intend to record the behavior of the all fireflies to recognize the weak ones, and enables them to update their locations by jumping to new locations in order to obtain find the solution, when the fireflies modified their locations that lead to modify the locations of whole population. The jumping operation increase probability of finding the optimal solutions, as so as increasing the performance of the proposed algorithm.

El-Sawy A. A. and et. al. [7] suggested a new approach that combines between two optimization techniques, "ACO and FFA". The propose approach was tested on many optimization problem such as benchmark problems, by applying this combining approach the researchers found that his performance was better than the performance of each approach when it is work alone.

Garsva G. and Danenas P. [10], their paper suggests new approach for linear classifier optimization method. Experimental results refer to the ability of proposed approach get competitive or better results compared to another similar approach. The linear classifier optimization approach can used to solve several classification problems with efficient solutions.

Asokan K. and Ashok Kumar R. [2], they propose an innovative optimization approach for defining bidding techniques, is shown as a stochastic optimization problem. The firefly algorithm introduced to this problem to optimize the search operation for best solution. By applying this approach the GENCOs profit maximizes in an effective way. Six suppliers was introduced to illustrate the main features of this approach, all results were displayed.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. FIRFLY ALGORITHM

In 2007, Firefly algorithm was used for the first time, [13]. It was used to optimize the Intelligence swarm algorithms. The method of this algorithm depends on the nature behavior of the firefly and the bioluminescent method for interaction between them [13] [15].

The difference in the value of light intensity is the value that is relied on by an objective function of an optimization problem. Depending on this value, fireflies update their locations as they move to the most attractive locations to reach the optimal solutions. Thus, light intensity that is related to the objective function is the characteristic of all fireflies [6].

A. Characteristics of Firefly Algorithm

Three basic rules were found for the Firefly algorithm, which rely on the main flashing characteristics of the behavior of living fireflies in nature. The rules were as follows:

1. All fireflies are "unisex" so fireflies will attract individual firefly.
2. The attractiveness is proportional to their brightness. The brighter fireflies attract other fireflies which has less bright. However, when the distance between two fireflies increase, the intensity should decrease
3. Fireflies move randomly, if the fireflies have a same brightness level.

By computing the value of the objective function, the firefly's brightness can be determined [16].

B. Functions of Firefly Algorithm

1. Attractiveness

Firefly's attractiveness function has its own form that can be illustrated in the decreasing function that described in equation 1.

"r is the distance between any two fireflies, r=0 is the initial attractiveness at r=0, and γ is an absorption coefficient which controls the decrease of the light intensity" [2].

$$\beta(r) = \beta_{0\, e^{(-\gamma r^m)}} \qquad with\, m \geq 1 \qquad (1)$$

2. Distance

If there are two fireflies i and j, the distance between them can be found in the following equation:

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \qquad (2)$$

"x_i, is the k-th element of the i -th firefly position within the search space, and d denotes the dimensionality of a problem" [9].

3. Movement

The following equation shows the movement of fireflies attracted by the most attractive fireflies [3]:

$$X_{i+1} = x_i + \beta_0 e^{(-\gamma r)2} (x_i - x_j)^2 + \alpha\, (rand - 0.5) \qquad (3)$$

"The second term is due to the attraction while the third term is the randomization with being the randomization parameter". Where "rand" was a random number generator, "rand" value was distributing in the range of [0, 1] [8].

The following pseudo-code form presents the firefly algorithm [6].

1. Algorithm's parameters initialization:

- Number of fireflies (n_f).
- $\beta 0$, γ, α
- Maximum number of generations (iterations, n_{itre}).

2. Define the objective function f(x), x = (x1. . ., xd)^T.

3. Generate initial population of fireflies xi (i = 1, 2 . . ., n). Light intensity of firefly I_i at x_i is determined by value of objective function f(x_i).

4. While k<n_{itre}

5. For i = 1:n

6. For j = 1:i

7. If (Ij> Ii) move firefly i towards firefly j in d-dimension according to Eq. (3); End if.

8. Obtain attractiveness, which varies with distance r according to Eq. (1).

9. Find new solutions and update light intensity

10. End for j.

11. End for i.

12. Rank the fireflies and find the current best

13. End while

14. Find the firefly with the highest light intensity.

The following equation represents the initial population of fireflies :

$$xi = LB + rand \cdot (UB - LB) \qquad (4)$$

Where LB and UB denotes the lower and the upper bounds of i-th firefly [6].

IV. ARDUINO MEGA2560

Arduino "is an open-source physical computing platform based on a simple I /O board", It takes the inputs of variety sensors or switches, and has the ability to control many devices and send different types of outputs such as lights, and other outputs, Arduino is therefore used to develop objects that need to interact with their external environment or with other objects, as they can interact with computer programs such as flash and processing [17]. However, Arduino can accomplish many projects on its own; and it has a special development environment for writing programs [5]. We use Arduino Mega 2650 to develop our system. See figure 1.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

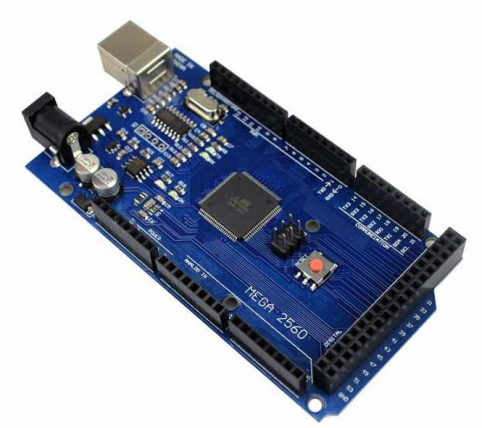


Figure 1: Arduino mega2560

Arduino designed for people with little technical and programming expertise, the use of Arduino allows these people to create a sophisticated model for project design and interactive artworks. People who have a strong technical background will be very easy for them to apply first steps with Arduino [14].

V. GLCD 192*64

A graphic LCD "liquid crystal display" is one of the electronic technologies that used in visual display, and also used in different gadgets and information output sources.

Through precise electronic signals, GLCD technology can employ manipulating tiny crystals of a contained liquid crystal solution to perform graphic display operations over a two dimensional screen.

LCD technology uses electron firing gun to produce a pixel based display over monitor screens, if traditional CRT "cathode ray tube" technology is compared with LCD technology; the latest technology is more successful [11] [20].

Figure 2: Graphic LCD.

We use Unified Modeling Language (UML) to develop our proposed system; use cases diagram, activity diagram and sequence diagram were used to analyze the system. See figures 3, 4 and 5.

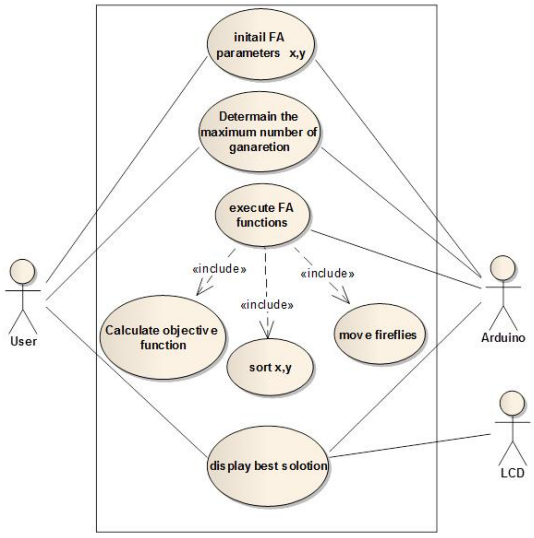


Figure 3: Use case diagram.

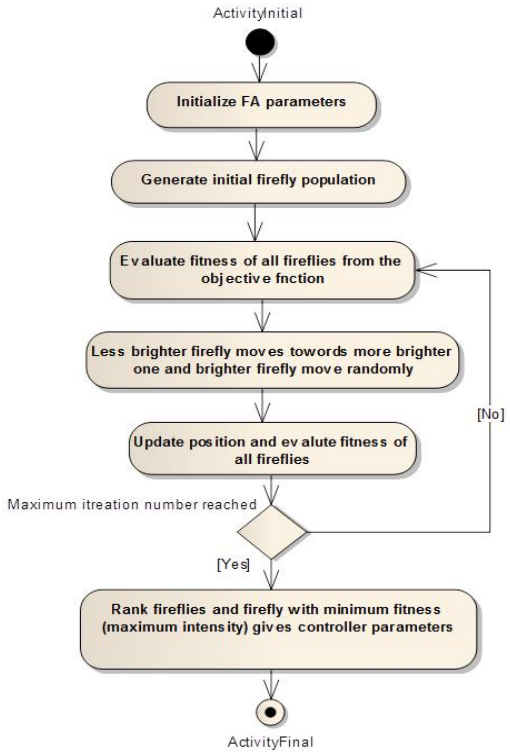


Figure 4: Activity diagram for firefly algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

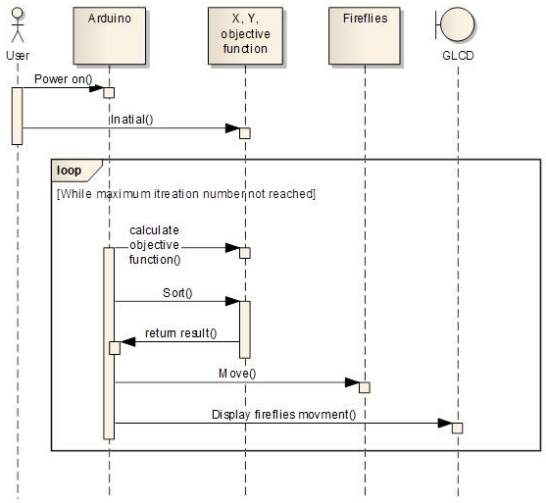


Figure 5: Sequence diagram.

VI. OPTIMIZATION TO FIND THE MINIMUM AND MAXIMUM VALUE OF VARIABLES IN MATHEMATICAL EQUATIONS USING FIREFLY ALGORITHM.

The firefly algorithm used to solve many optimization problems. For doing this we need to determine the objective function and the control parameters that can be decision variables, is given in equation 5:

$$\min, max \rightarrow f(x,y) = x + cy^2 - xy\,, (x,y) \in (-w,w)$$

Where c denote "any constant number", and w denote the "lower and the upper bounds of i-th firefly", the value of the x and y variables is choose randomly.

In our research we implemented firefly algorithm based on Arduino under windows XP or windows seven operating system and the result displayed at GLCD. In the implementation of any technique of metaheuristic techniques, the control parameters must be initialized. This also applies to the firefly algorithm, and it is very important to choose appropriate values for the control parameters to find the best solutions, the assigned values of control parameters are determining the performance of this method. Our selection of these parameters depends on a wide range of experimental results.

The control parameters displayed as following:

A. n_f: is the fireflies number, in all examples $n_f = 45$. We choose this value because when we set up n_f to a large number "more than 100 fireflies", the results of our experiments are not change greatly and the execution time was increase with no improvement at all.

B. n_{iter}: is iterations number, n_{iter} is another control parameter of the firefly algorithm which must be appointed to execute the algorithm until achieving the convergence of the minimization of the error. In order to find the global optima, the firefly algorithm was not need to large number of n_{iter}.

In all our experiments, n_{iter}= 50. We found that the value was a suitable,

When we increase the n_{iter} more than 50 iteration, the result does not improve.

C. β_0: The initialization value of attractiveness, as several suggestion for many optimization problems the value of β_0= 0.1. In the present study we take an above value, which give to us very good results.

D. γ: is the absorption coefficient, where $\gamma = 1$ in our paper, this value produce a convergence of the algorithm quickly.

E. μ: the value of potential coefficient, it can be assigned to any positive number. The value of μ = 0.1in our study.

F. α: the value of randomization parameter. This control parameter can be any number on the interval [-2.048, 2.048]. The randomization degree was determined by α value. The parameter α was so important because it was allowing to produce a new solutions, so as not to stuck in a local minimum. In our research α = 0.1, we choose this value in order to avoiding perturbations on the firefly.

First we must choose control parameters value. The firefly algorithm is performed iteratively until reached the number of iterations. To remove the stochastic effect and avoid premature convergence, 20 independent executions have been carried out. Then, the firefly with the best fitness value was selected as the optimal solution to the given problem.

VII. EXPERIMENTAL RESULTS

In this section we check the performance of our work, it has been tested with a large collection of examples, and the results were excellent in all cases. In this section we consider only one of these examples. These examples were selected to illustrate the variety of situations that could be applied using this method

The example in this paper is shown in Figures 1, 2 and 3. Three different figures are displayed: on the Figure 1, we show the first iteration that display the primary locations of the fireflies, on the figure 2,we show the eight iteration that display the new locations of the fireflies and on the figure 3, we show the twenty one iteration that display the fireflies are reaching to goal. As we say before our results will display on GLCD, this information is: the number of the iteration (Itre), the minimum value (x) and the maximum value (y) of variables in mathematical equations, the value of lightness (l), and finally the value of error (E).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 6: The first iteration that display the initial locations of the fireflies

Figure 7: The ninth iteration

Figure 8: Twenty one iteration that display the fireflies are

VIII. CONCLUSION

The firefly algorithm is an effective technique in solving global optimization problems. The firefly algorithm was used in this paper to find the minimum and maximum value of variables in mathematical equations using Arduino microcontroller.

The suggested approach depends on choosing the value of control parameters of firefly algorithm like: number of iterations, absorption coefficient, determination of the objective function, and population number of fireflies. Experimental results show that the results that were obtained is matching with desired results.

It is therefore possible to say that swarm Intelligence algorithms are highly efficient in solving optimization problems, including finding minimum and maximum value of variables in mathematical equations.

REFERENCES

[1] Abdullah A. and his colleagues, (2012), "A New Hybrid Firefly Algorithm for Complex and Nonlinear Problem", Springer, distributed computing and artificial intelligence, AISC 151, pp 1.

[2] Asokan K. and Ashok Kumar R. ,(2014), "Application of Firefly algorithm for solving Strategic bidding to maximize the Profit of IPPs in Electricity Market with Risk constraints", international journal of current engineering and technology, vol.4, no.1, pp 37-44.

[3] Benabid R. and his colleagues, (2014),"Application of Firefly Algorithm for Optimal Directional Over current Relays Coordination in the Presence of IFCL", I.J. intelligent systems and applications, pp 48.

[4] Bidar M. and Kanan H. R., (2013), "Jumper Firefly Algorithm", international conference on computer and knowledge engineering ICCKE , Qazvin, Iran, pp 278-282.

[5] Cytron technologies, (2011), "Fun and learning with Arduino projects ", Beginner guide, pp 4.

[6] Dhillon M. and Goyal S.,(2013), "PAPR Reduction in Multicarrier Modulations Using Firefly Algorithm", international journal of innovative research in computer and communication engineering vol. 1, issue 5, July , pp 1272.

[7] El-Sawy A. A. and his colleagues, (2012), "A Novel Hybrid Ant Colony Optimization and Firefly Algorithm for Solving Constrained Engineering Design Problems", journal of natural sciences and mathematics, qassim university, vol. 6, no. 1, pp 1-22.

[8] Farook S. and Raju P. S., (2014), "Metaheuristic Algorithms for Capacitor Sitingand Sizing to Improve Voltage Profile", International Electrical Engineering Journal (IEEJ), Vol. 5, No. 1, pp. 1211.

[9] Fister I. and his colleagues, (2014), "On the Randomized Firefly Algorithm", Springer International Publishing Switzerland, pp 34.

[10] Garsva G. and Danenas P., (2014), "Particle swarm optimization for linear support vector machines based classifier selection, Vilnius University, vol. 19, no. 1, pp 26–42.

[11] Jackson K. (2010), "Using a HD44780 Compatible LCD module with an Arduino microcontroller and Programming Environment", https://www.arduino.cc/en/Tutorial/LiquidCrystalDisplay.

[12] Kwiecien J.andFilipowicz B.,(2012), "Firefly algorithm in optimization of queuing systems", bulletin of the polish academy of sciences, technical sciences , vol. 60, no. 2, pp 1.

[13] Sarbazfard S. and Jafarian A., (2016), "A Hybrid Algorithm Based on Firefly Algorithm and Differential Evolution for Global Optimization",

[14] International journal of advanced computer science and applications, vol. 7, no. 6, pp 96.

[15] Schmidt M., (2011), "Arduino A Quick-Start Guide", North Carolina, Texas, Pragmatic Programmers, LLC, Second Edition.

[16] Wang G. and his colleagues, (2012), "A Modified Firefly Algorithm for UCAV Path Planning", international journal of hybrid information technology, vol. 5 issue 3, p123.

[17] Winkler F., (2007), "Arduino Workshop", pring,http://www.arduino.cc.

[18] Yang X. S., (2009), "Firefly algorithms for multimodal optimization", lecture notes in computer sciences 5792, pp 169.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[19] Yang X.S.,(2010), "Firefly Algorithm, L´evy Flights and Global Optimization", Springer Verlag London Limited, pp 1.
[20] www.vishay.com/docs/37364/37364.pdf.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Analysis of android bugs for mobile applications

Tawfeeq M. Tawfeeq Al-Flaih
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
Tawfeek.mflaih@yahoo.com

***Abstract*— Open source Mobile applications have gained a lot of popularity in today's world. But most of these mobile applications tend to be buggy which may affect the user experience and thus they need quick bug fixes. For our research we have taken into account 10 applications from different domains. The aim is to study the bug reports of these applications and analyze them. Our objective for this research is to understand the life cycle of Android bugs and the relationship between the various domains and ratings with the number of bugs.**

Keywords: mobile applications, Android bug report, Google play store, bug fixing, bug report quality

I. INTRODUCTION

Mobile devices have become an important part of people's lives in recent years. Smartphone's have gone beyond their basic communication functions and now offer many features that in the past belonged solely to the domain of personal computers. As a result, companies have developed mobile versions of applications that were originally for other platforms. There is also a large quantity of applications developed specifically for mobile [13]. Multiple Apps stores were created by large companies to accommodate and manage their platforms Apps. On the other hand, and due to the wide spread of these mobile application, software repositories were used to maintain and share open source code for such applications. Software repositories such as source control, bug and, communication repositories are widely used in large software projects [7].

Application stores (e.g., Google Play, Apple App Store and BlackBerry App World.) have changed the traditional software development concept by providing their own platform for the rapid growth of mobile apps .in the past few years, mobile apps have exploded into a multi-billion dollar market and their popularity become hugely wide among consumers and developers. Mobile app downloads have risen from 7 billion in 2009 to more than 197 billion in 2017. In the same time, mobile apps numbers have also increased: Google Play now hosts over 28 million mobile apps [2][3].

In this paper we want to shed some light to understand the life cycle of open-source Android Apps bugs. To accomplish our goal we analyzed the bug reports of ten open-source Android Applications trying to understand the life cycle of these bugs. Furthermore, we are trying to measure the quality of these bug reports.

II. MOTIVATION

Recently, Android platform and its applications have gained tremendous popularity. The septal to entry in applications development and deployment has drop, due to easy distribution across application stores such as Apple App Store [8]. This means that's apps and app updates are subject to limited audit before deployment, and in this case there are many error-prone applications in the market and affecting user experience. Open source Mobile applications have gained a lot of popularity in today's world. But most of these mobile applications tend to be buggy which may affect the user experience and thus they need quick bug fixes. Most of open source Mobile applications have the bug report to gain feedback from users. User reports bug he have and describes some bug information. The bug will be opened and finally be closed. From "open' to "close", there is a life cycle of bug. Understanding the life cycle of bug can help us to reduce the bug occurrence.

Compared with iOS application, the Android applications will be run on many kinds of mobile drives. In this case, the cost of checking and fixing bugs for Android application will be more expensive than on iOS. For reducing this cost, the most effective method is that decreasing the number of bugs before the application released. We hope we can find some properties about Android application through analyzing the bugs and bug report.

III. RELATED WORK

There is a lot of research performed related to life cycle of bugs in the android applications. Bhattacharya P. et. al. in the paper "An Empirical Analysis of Bug reports and Bug fixing in open source android apps" performed an empirical analysis so as to understand the bug fixing process in the Android platform and Android based applications. In order to perform their research, they selected 24 popular android applications. They selected apps depending on certain metrics and analyzed the bug fix processes. This included the bug fix time, bug categories, bug priorities and also the interest of the users and developers to fix the bugs. On comparing the life cycle of bugs on Google Code and Bugzilla they found that lack of certain bug report attributes affects the bug fix process. They investigated the categories of security bugs in Android applications. On conduction of the analysis, they found that even though the contributor activity in the projects is high, the involvement of developers is less. Also, triaging bugs is still a problem even though the bug reports are of high quality. They

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

observed that the non security bugs required less time to fix even though the quality of security bug reports was better.

The MSR challenge provides platform for the researchers to add their mining tools and approaches to the challenge. There is a research done in the android platform for analyzing bugs and finding some interesting facts of those reports. This research is performed by Shihab E. et al. in the paper "Mining Challenge 2012: The Android Platform". The work is performed on the change data and bug report data of the android platform which has been extracted from GIT repository and android bug tracker. They selected sub-projects for change data from the android those are Kernel/linux, kernel/omap, kernel/tegra, Kernel/Samsung, kernel/qcmu, kernel/experim-enttal, platform/frameworks/base, platform/external, Bluetooth/ bluez. In the change data analysis the result states that the numbers of authors are more than the number of committers, which shows GIT has fewer contributors than committers so as to fix the issues. In a similar way for the bug report they selected 10 different components. Those are Market, Docs and Build, User, Web and System, GfxMedia, device, media, Google dalvik, tools, applications, platform and no component. The result set the average fix time for bug found is 2.34 months, most of bugs were not assigned to any of the particular component, the committers commit on the bug report only once during the project and 99% of the bugs are of medium priority and also it has a length of average 189 words.

Syer et al. performed a study on comparing the mobile applications with different desktop application. They considered two aspects for comparison, the size of the code base and the time to fix the defects. For conducting their study they considered 15 popular open source android applications and 5 different desktop applications. They found that there is a large difference between the mobile apps and desktop apps in some respects, while in some respects they are similar. They found that the core developers in mobile apps are very small as compared to desktop applications. Thus it is necessary to pay attention to mobile development now by keeping aside the desktop applications. In our research, we are going to study the life cycle of bugs in open source android applications.

IV. RESEARCH QUESTIONS

In order to conduct our study, we have identified the following research questions. Our objective of this research is to answer these questions:

1. How can the quality of bug reports help the contributors fix the bugs sooner?

2. What is the relation between domain of Application and number of Bugs?

3. What is the relation between the rating of Application and number of Bugs? Our methodology to answer these research questions is described in the next section.

V. METHODOLOGY-STUDY DATA

A. Selection Criteria:

The mobile applications we selected in this project are open source applications. There are millions of mobile applications in the market today. However, only 10 mobile applications are needed. So the selection criteria are narrowed by selecting android applications. The reason is its number of options available to select an application, the popularity of applications people using, these applications are available free of cost and main important reason is availability of its own bug repository with some of its applications.

The android mobile applications are downloaded from Google Play store. There are two categories of applications available in the play store. Those are free and paid. As name implies the applications can be downloaded at free of cost and paid applications can be downloaded by paying for it. The advantage of play store is, it provides 26 categories to choose the application. Moreover, the play store provides a detail like category, number of downloads, number of people rated it and also some time it provides link to Git repository. It is important to remember that not all the free applications of the play store comes with the GIT repository.

The Git repository provides all the details necessary for the bugs to analyze. The complete life cycle of the bug can be observed. The bugs from the initial release to the present releases can be found. The open bug count, closed count, data and time they were reported and fixed, and contributors and commenter's details can also be studied from here.

The table 1 provides the details of the 10 mobile applications selected for the project. The details are its category, the number of downloads, number of people rated and wrote the review, total number of releases and the bug count which is sum of the open bugs and closed bugs from the first release of the application.

We use some tools as CUEZILLA tool to measures the quality of new bug reports [4].

TABLE I. APPLICATIONS DATA

Name	Category	Downloads	Ratings	Releases	Bugs Count
Zxing: Barcode Scanner	Utility	100,000,000 – 500,000,000	704060	16	372
FBReaderJ	Education	10,000,000 – 50,000,000	128429	306	325
Wordpress	Editor	1,000,000 - 5,000,000	63185	69	131
Keepassdroid	Security	1,000,000 - 5,000,000	28904	110	317
Ifixit	Utility	500,000 - 1,000,000	5760	28	251
Simon Tatham's Puzzles	Game	100,000 - 500,000	31469	56	230
Car Cast	Multimedia	100,000 - 500,000	1240	77	121
BetterBattery Stats	System	100,000 - 500,000	7,986	139	601
AnkiDroid	Education	1,000,000 - 5,000,000	18,166	389	745
XBMC	Multimedia	100,000 - 500,000	281	81	343

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

RQ1: How can the quality of bug reports help the contributors fix the bugs sooner?

Our first research question was to understand how the quality of bug reports will help the contributors and developers to fix the bugs easily and quickly. In order to answer this question, we have taken into account different characteristics of bug reports. These characteristics include the length of description of the bug in the bug reports and number of keywords found in the description [1][5]. They keywords that we have considered are version, component, security, vulnerability, attack, failure, error, crash, buffer overflow, buffer overrun, question, problem, invalid, and incorrect. For every bug in the bug report for each application, used a script to find out the length of the description of the bug. Also we wrote a script which took the input as the above mentioned keywords and found them in the bug descriptions. The descriptions having highest number of keywords along with sufficient description length were chosen. The bug descriptions which were too lengthy and did not have high count of the keywords were ignored. Also the bug descriptions which were very short in length but had large number of keywords were discarded. Further, we calculated an average of description length and the average of number of keywords for each application.

The table 2 below shows an example of how each bug from every application was analyzed to find the description and number of keywords and the corresponding time spend to fix the bug.

TABLE II. BUG REPORT QUALITY

App	Bugs ID	Bug Title	Start Date	End Date	Time	Length of Description	Number of Keywords
AnkiDroid	105	Allow users to change AnkiDroid directory if current one is invalid	2/3/2015	2/15/2015	12D	190	3
FBReaderJ	219	Fatal exception in BookDownloaderService	1/10/2013	Open	N/A	171	3
Zxing	308	Possible ReedSolomon deoding problem	2/18/2014	2/20/2014	2D	253	4
CarCast	71	Review if debug mode is needed on release builds	10/26/2012	7/1/2014	978D	30	3
ifixitAndroid	106	SSL errors on Android 2.2	9/9/2013	9/10/2013	1D	313	4
KeepassDroid	39	2nd try…	10/21/2010	10/24/2010	3D	68	1
sgtpuzzles	9	Build Failed	2/19/2015	Open	N/A	79	3
WordPress	104	Bugfix - iploading post thumbanails	3/16/2013	3/16/2013	3min	86	2

Result:

We obtained the following results as described in the table 3 below. On careful analysis, we found that the time span required to fix the bugs was less for the bugs which had good description length along with large number of keywords. The keywords and description made it easy for the developers and contributors to understand the bugs and get them fixed as early as possible. As we can see for Zxing application the average length is 60 and average number of keywords for the bug report is 112 so the average fix time is less 10 days. Similarly for Simon puzzle application the average length is 46 and in proportion to the length, average keywords is 40 so time span is 6 days.

TABLE III. AVERAGE LENGTH, AVERAGE NO. OF KEYWORD AND NO. OF DAYS FOR BUGS IN EACH APPLICATION

Name	Total No of bugs	Avg Length	Avg Keywords No.	Avg Time Span in days
Zxing: Barcode Scanner	372	60	112	10
FBReaderJ	325	48	52	16
Wordpress	131	18	16	99
Keepassdroid	317	36	20	59
Ifixit	251	57	26	112
Simon Tatham's Puzzles	230	46	40	6
Car Cast	121	43	6	139
BetterBatteryStats	601	55	87	15
AnkiDroid	745	42	54	3
XBMC	343	52	61	4

RQ2: What is the relation between domain of Application and number of Bugs?

The second research question is focus on the relation between domain of application and the number of bugs. In our research, we choose four domains as our research objectives. We select four open source applications for each domain. In the same time, all of the applications we choose have substantially the same ratings. That means those applications have the same evaluation. Then we calculate the bug count for each application, and compare them based on the same domain.

Result:

The charts below show the relation between the application domain and the number of the bugs. As shown in the figures (1) are no clear variation between results number of bugs are very close between the domains. One of the apps in the education domain (ankidroid) has more bugs in comparisons with other domain but this is may not be because the domain bug may because the application itself or the nature of the team who developed this application.

RQ3: What is the relation between the rating of Application and number of Bugs?

The last research question to answer is relation between the number of ratings and bug count. In Google play store the people who downloaded it rate the application from 1 to 5 stars that is, from average to very good. Along with

ratings of the stars, the people write their review on the application use. The rating is sum of the people rated application and people wrote reviews for it. The bugs count is from the GIT repository. The analysis is made for each selected 10 applications.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

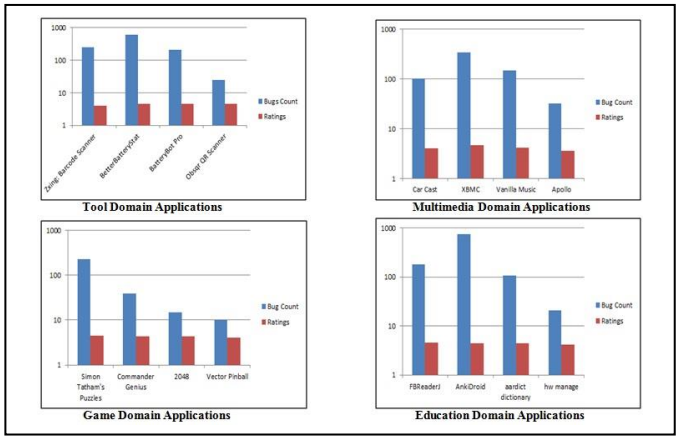


Figure 1. Graphs for rating of each domain

Result:

The results is found by analyzing the figure (2) below, which shows the relation between the ratings of the application and bugs count for each individual application. As we can see from figure 2, the rating and bugs count are inversely proportional. Higher the rating for an application lower is the bug count. In 10 applications, all the 9 application supports the conclusion other than one application XBMC, which is a multimedia application similar to VLC player. The reason for this is the number of downloads 100,000 - 500,000. This Application also has very less feature compared to other media players which makes its less popular and more buggy. This download number reveal that the people are less interested in using it so the less number of ratings 271. Hence the greater number of bugs counts.

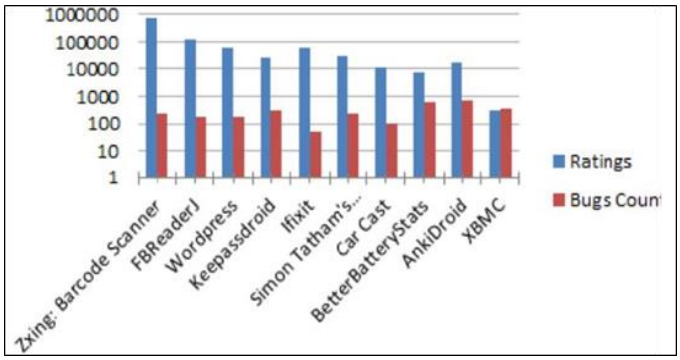


Figure 2. Relation between rating of application and numbers of Bugs

B. *ZXING Analysis:*

During the initial phases of the data collecting and analysis, ZXING application raised many question because of its high download and rating numbers. Also, we notice that its bugs count relatively low. Therefore, we decided to go further analyzing this application, and seek some answers and explanation for these numbers. To accomplish that goal, we extracted the end users reviews from ZXING application page in Google play store and classify them into feature requests and bug reports. Then, we tried to see if the users are satisfied with application and not asking for many features.

Analysis Approach:

After extracting the reviews from the Google Play Store, to answer our questions we needed to classify them in feature requests and bug reports. To achieve this, we wrote an algorithm that splits a text into sentences, normalizes them, and compares them with a set of linguistic rules to find if it matches any of them. We used two set of rules for our classification. One is based on the linguistic rules defined by Iacob et al. for feature requests [9], the other one is based on the linguistic rules they defined for bug reports [10]. We adapted the syntax of the rules to work with OpenNLP [12], the API we used for part-of-speech tagging. To classify issues, we also modified some rules, because the way people talk when reporting an issue is somewhat different than the way they talk when leaving a review at an app store. Table (4) shows some examples of linguistic rules for identifying feature requests and an example text that would be a match for each. Table (5) shows some examples of linguistic rules for identifying bug reports and an example text that would be a match for each.

TABLE IV. EXAMPLES OF RULES TO IDENTIFY FEATURE REQUESTS

Rule	Text match
Would be <adjective> if	It would be great if
Would <adverb> like to <verb>	Would really like to see
Needs option to	Needs options to share posts

TABLE V. EXAMPLES OF RULES TO IDENTIFY BUG REPORT

Rule	Text match
<adverb> annoying	Incredibly annoying
Won't <verb>	Files won't open
Keeps on crashing	Reader keeps on crashing

To create an algorithm that classified the reviews based on those rules, we used Lingpipe and Opennlp. We started by splitting the review in sentences, using Lingpipe [11] to recognize end of sentence tokens. Lingpipe is a toolkit for processing text using computational linguistics. After the review was split into sentences, we normalized each sentence, replacing common misspelled words and abbreviations. After, we used OpenNLP [12] to tag the sentence. OpenNLP is a machine learning based toolkit for the processing of natural language text. We used it to tag each word in the review sentence as a part of speech, e.g. for the text "it would be great"

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the tagger would generate "<personal_ pronoun> <modal> <verb> <adjective>".

We ran the algorithm twice for each review we extracted. The first time it compares the review sentence with the linguistic rules defined for feature requests, and if it matched one of the rules, it classified the sentence as a feature request. If a sentence of a review, then the review is marked as not feature request. The second time it did the same, but instead of comparing the sentence with the rules for feature requests, it compared them to the rules for bug reports. After the reviews in the database were classified, we counted the bug report and feature requests for the applications.

VI. DISCUSSION

Base on the results of three questions, we can clearly find out the quality of bug report affect the speed of fixing the bug. The keywords can make developers easily to understand and locate the bug. The length of bug report and the time for fixing is a negative correlation. This is easy to understand that the more information the developers get, the faster the bug can be fixed.

Also, the differences of domain generally have relation with quality of mobile application. Base on the same rating for each domain, the game mobile application have less number of bugs than other three domains. The one reason we think is that user have more patience for other three domains than for game. That means the mobile game developers have to pay more attention to reduce the incidence of bug. Through analyzing the rating of application and the number of bug, we found out the applications having higher ratings have less bug count. The bug can reduce user experience, thereby decrease the ratings.

VII. THREATS TO VALIDITY

The research would have been given better results if more number of applications were taken into consideration. For this study we have considered only 10 applications which is a very limited number. Also for our research question 2, we considered only 4 applications for each domain. If more number of applications were considered we could have got different results. In Addition, all the applications are written in the same language (Java). We did not explores the code of these apps such as the number of classes, the number of developers, and the experience of developers of these application, to get better results we supposed to select the apps that are close in the number of line of codes or the number of developers so that the domain will not be affected by the code.

When testing our review classification algorithm for Zxing analysis, we used our personal judgment to decide if the algorithm was right or not for identifying a feature request or a bug report. Therefore, the accuracy measures of the algorithm are biased.

VIII. CONCLUSION

In this paper, we analyzed some open source mobile applications, and get the relations between domain, rating, and the quality of bug report. Through analyzing the life cycle of bug, we realize that the users' behavior also affect the quality of mobile application. In this case, it proved that the importance of bug report. We also realize that the difference of user's patience for different domain. This also decides that there are different decisions of testing and fixing bug in the different domains. For the future work, we can increase the number of mobile applications which are analyzed to get richer data set. In the same time, we want to make clear for the correlation between length of bug report and the time of fixing. Is there a crest in their correlation? In other words, does the too much information in bug report effects the developers' understanding for the bug? We hope our research can provide some enlighten for who also analyze this area.

IX. REFERENCES

[1] Bhattacharya, P., Ulanova, L., Neamtiu, I., & Koduru, S. C. (2013, March). "An empirical analysis of bug reports and bug fixing in open source android apps". In Software Maintenance and Reengineering (CSMR), 2013 17th European Conference on (pp. 133-143). IEEE.

[2] Shihab, E., Kamei, Y., & Bhattacharya, P. (2012, June). "Mining challenge 2012: The android platform". In Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (pp. 112-115). IEEE Press.

[3] Syer, M. D., Nagappan, M., Hassan, A. E., & Adams, B. (2013, November). "Revisiting prior empirical findings for mobile apps: an empirical case study on the 15 most popular open-source Android apps". In Conf. Center for Advanced Studies on Collaborative Research (CASCON 13), IBM, 2013, pp. 283–297.

[4] Bettenburg, N., Just, S., Schröter, A., Weiss, C., Premraj, R., & Zimmermann, T. (2008, November). "What makes a good bug report?". In Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering (pp. 308-318). ACM.

[5] Asaduzzaman, M., Bullock, M. C., Roy, C. K., & Schneider, K. A. (2012, June). "Bug introducing changes: A case study with Android". In Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (pp. 116-119). IEEE Press.

[6] Martie, L., Palepu, V. K., Sajnani, H., & Lopes, C. (2012, June)."Trendy bugs: Topic trends in the android bug reports". In Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (pp. 120-123). IEEE Press.

[7] Jongyindee, A., Ohira, M., Ihara, A., & Matsumoto, K, I. (2011, November). "Good or Bad Committers? A Case Study of Committers' Cautiousness and the Consequences on the Bug Fixing Process in the Eclipse Project". In Software Measurement, 2011 Joint Conference of the 21st Int'l Workshop on and 6th Int'l Conference on Software Process and Product Measurement (IWSMMENSURA) (pp. 116-125). IEEE.

[8] Sinha, V. S., Mani, S., & Gupta, M. (2012, June). "Mince: Mining change history of android project". In Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on (pp. 132-135). IEEE.

[9] Iacob, C., and Harrison, R. (2013, May) "Retrieving and analyzing mobile apps feature requests from online reviews". in Proceedings of the 10th Working Conference on Mining Software Repositories, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, pp. 41–44.

[10] Iacob, C., Harrison, R., and Faily, S. (2014) "Online reviews as first class artifacts in mobile app development" in Mobile Computing, Applications, and Services. Springer, pp. 47–53.

[11] alias i, LingPipe, 2003U″ 2011. [Online]. Available: http://alias-i.com/lingpipe/

[12] T. A. S. Foundation, OpenNLP, 2010. [Online]. Available: https://opennlp.apache.org/

[13] Jiang, H., Yang, H., Qin, S., Su, Z., Zhang, J., Yan, J.,(2017, Dec) "Detecting Energy Bugs in Android Apps Using Static Analysis" In: Duan Z., Ong L. (eds.) Formal Methods and Software Engineering. ICFEM 2017. Lecture Notes in Computer Science; Vol. 10610: pp 192-208 : DOI: 10.1007/978-3-319-68690-5_12.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Enhanced Pipeline for Improving Reliability using Evolution and Optimization Technique

Dr Narasimha Rao Yamarthi[1], Mr Adinew Belay[2], Mr Abdisa Lechisa[3], P.Narasimha Rao[4].
Professor, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
HOD, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
narasimha.yamarthi@gmail.com, adinewb2@gmail.com, abdisalechisa@gmail.com,pnr555@gmail.com

Abstract

Objective: In advanced digital systems the propagation delay plays a vital role to optimize the performance of an individual processing element. In the present paper the advantages and flaws of various pipelines are discussed. In the present paper the performance evaluation of different pipelines is done in terms of various parameters like timing delay, throughput, and average delay. These factors are very important in achieving parallel computing in fast processors.

Methods/Analysis: In the present paper the proposed pipeline is compared with Traditional Pipeline, Wave Pipeline, and mesochronous Pipeline. In all the cases the throughput, Timing delay, and average time delay are compared and proved that the proposed method has produced more effective parameters. All the observations are made at 4-stage pipeline. The design analysis is done with the simulation software Proteus. The accurate data wave reliability is tested in Proteus. The propagation delay is illustrated with the help of Electronic Work Bench. The readings are distinguished at different data frequency rates like 100, 500 and 1000MHz.

Findings: It is observed that a four stage proposed pipeline has good throughput when compared with other pipeline clock schemes. Based on the observations the wave pipeline is superior to any other method in terms of through put and data reliability. The proposed method achieved slight improvement when compared with wave pipeline. The data reliability is good in proposed method at different frequency stages.

Novelty/Improvements: To achieve parallelism in advanced processors, pipeline technique is the best method proposed by many architectural designers. In real time operating systems the pipeline helps in message passing and fetching in due time. But there are many design and operational factors need to be considered in achieving high performance. The Propagation delay is one of the important factor need to consider in pipeline design.

Keywords: Parallelism, Pipeline, Clock Scheme, Propagation Delay, Efficiency, Throughput, Reliability

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

1. Introduction

In the present paper a data pipeline is discussed in terms of its performance based upon the clock scheme. There are different clock schemes effectively acting on internal latches of pipeline. Different timing constraints are indicated in the previous work held in the present area[1,2]. The digital pipeline play very important role in processing the data with memory and I/O devices. Pipelines plays most vital role to avoid bottle neck. Pipelined processors can be clocked a fast clock rate and thus can have reduced cycle times (more cycles/second by a fast clock) than un-pipelined implementations of the same processor[3]. In traditional pipeline the flow of data that is inputting the data, intermediate process, and outputting the data from stage to stage is controlled by common clock cycle. The process between all stages is controlled by common clock cycle. The intermediate latches are used between stages to hold the intermediate process results. The stages are basic combinational circuit. Intermediate latches can also be used for delay balancing in data path[4].

In the present paper a linear pipeline with new clock scheme is discussed with multiple parameters which are set to prove the efficiency. For a general pipeline the time delay is denoted with τ.

Where, $\tau = \tau_m + \tau_l$ ---------- (1)

A linear pipeline with k stages uses k cycles to fill up the pipeline and n-1 cycles are needed to complete the remaining n-1 tasks. In the present design a static pipeline of unifunction is discussed with their performance. Here the pipeline is designed to operate at different pipeline bandwidths. The bandwidth represents the number of bits processed per unit time. Here the performance of the pipeline is measured in two factors[5].

Pipeline Efficiency: The efficiency of linear pipeline is measured by percentage of busy time-space spans over the total time-space spans over the total time-space span, which equals the sum of all busy and idle time-space spans [5]. Let *n*, *k*, and τbe the number of tasks, number of pipeline stages and the clock period of linear pipeline respectively, then the pipeline efficiency is defined by

$$\eta = \frac{n}{k + (n-1)} \quad \text{---------- (2)}$$

Throughput: The number of results that can be completed by a pipeline per unit time is called its throughput. This rate reflects the computing power of a pipeline. Throughput can be defined as shown,

$$w = \frac{n}{\tau} \quad \text{---------- (3)}$$

Average Delay in a stage is,

$$\tau_{avg} = \frac{T}{k} \quad \text{---------- (4)}$$

The time required to finish i^{th} instruction in a pipeline computer is T_i

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$T_i = (N_i + k - 1).\frac{T}{\kappa} \quad \text{---------- (5)}$$

Where,

τ_l= the delay of each interface latch
τ_m = the delay through longest logic path
k= the number of stages in a functional pipe
T= the total pipeline delay in one instruction execution
n= the number of instructions contained in a task
N_i= the length of vector operands used in the i^{th} instruction
W= the throughput of the pipeline computer
T_i=the time required to finish i^{th} instruction in a pipeline computer
η= the efficiency of a pipeline computer

In most of the cases the pipeline performance efficiency depends on the effective clock signalling. There are some mostly used clocking schemes are already playing an important role in steering the pipeline performance, such as synchronous, Asynchronous, Mesynchronous, and Plesynchronous clock schemes. In synchronous clocking maximum power consumption occurs due to global data. Higher clock speed is required and less clock periods will be used for computations. In Asynchronous clocking, the drawback is that the hardware and signalling overhead involved in the local communication and in any timing constrains that are required by particular choices of signalling protocols [6,7]. Plesynchronous interconnect only occurs in distributed systems like long distance communications. The data can be duplicated if the transmit frequency is slower than the receive frequency. These problems can be overcome with the new clock scheme. The above factors are observed by considering four different pipeline techniques, out of which one is the new method.

i) Conventional pipeline: In conventional pipeline system a single clock pulse is applied to manage the data transmission through the registers in the pipeline. But it will create a clock skew in thepipeline which will decrease the data speed from one stage together stage [8,9,10,11].

ii) Wave Pipeline: Smaller clock periods are achieved in wave pipelining[12] byreducing the maximum propagation delay (τ_m) by splitting the stages into number of stages[13]. The width of the clock pulse will be approximately equal to the difference between maximum and minimum propagation logic path delays between pipeline stages.

iii) Mesynchronous Pipelining: The propagation delay is reduced and the clock synchronization is controlled by introducing a delay element in the path of clock signal of Mesynchronous pipelining [14]. The delay element is almost equal to the logic path delay between pipeline stages.

2. Enhanced Method

A four stage pipeline is constructed to analyse pipeline operation and process the data. A four stage pipeline is proposed because; an *n* stage pipeline performs *n* faster operation in any type processor. Intermediate latches are used between stages to hold the intermediate process.

In order to achieve proper capturing of data at the output proper clock timing must be done between stages. The timing requirements must be met between clock and data edges at

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the inputs to the output. The clock period must be such that the output data is latched after the latest data has arrived at the outputs and before the earliest data from the next clock period arrives.

In the present method the logic gates at individual stages create simple delay in producing the clock to the next stage. Until the logic gates identify the next binary bit from previous stage it will not allow the clock generator to pass the next clock pulse to the next stage of the circuit as shown in Figure 1.

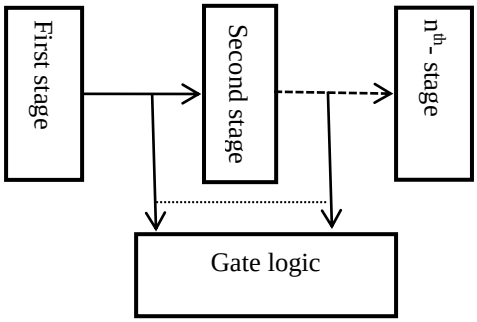


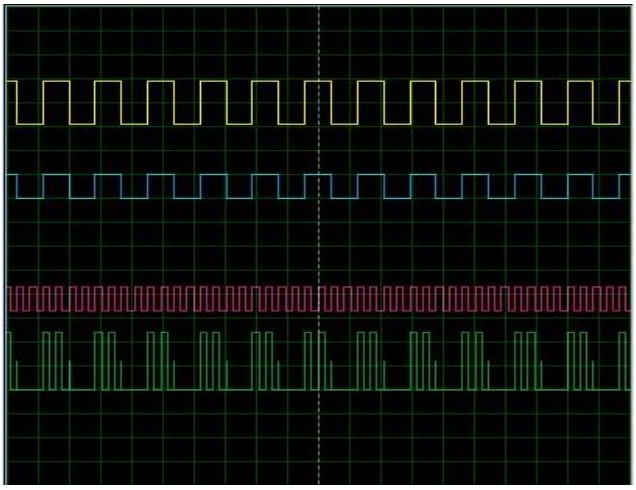
Figure 1. Proposed Method with new clock scheme

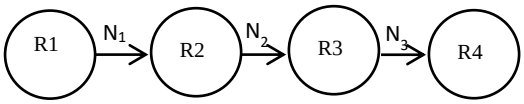
In the present paper this method is compared with other traditional, wave and mesochronous pipelines in terms of timing delay, throughput and average delay.

In traditional pipeline the data propagation is not accurate when compared with other optimized pipeline techniques. In synchronous pipeline a common clock pulse is unable to synchronize all stages at different frequencies.

In traditional pipeline the data propagation is not accurate when compared with other optimized pipeline techniques. In synchronous pipeline a common clock pulse is unable to synchronize all stages at different frequencies. It results loss of data bits in transmission. In asynchronous traditional pipeline as the stages increases the circuit complexity also increases and hence propagation delay. The propagation delay increases and results data latency in prior stages. Hence, the reliability is deprived in traditional pipeline. There may be a chance of fault event due to miss match between clock rate and data rate. This causes failures at individual phases [15]. If the failure rate is repeatedly occurring as shown in Figure 2 the reliability of the pipeline present system will be low. In Figure 2 the third and fourth pulses are input and output pulses respectively. In the present traditional pipeline system, there is a failure case identified at each third and fourth pulse due to fault occurrence because of synchronization problem between clock and data frequency. If the fault detection at any stage is N_k, where k is number of the stage, then the faults at individual stages can be represented with the help of model [16,] as shown in Figure 3. R_1 and R_2 are the reliability factors of stage 1 and stage 2 respectively, and reliability can be represented upto n stage like R_n.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500





In Wave pipeline and Mesochronous pipeline based system the reliability is improved when compared with traditional method. These two are the optimized clocking methods used to reduce the failures. In wave pipeline likewise faults arouse due to propagation difference between longest and minimum path difference between stages. This leads irregular data propagation failures as shown in Figure 4 some data loss is observed in fourth clock pulse. These failures are observed due to clock skews due to (D_{max}~D_{min}) of wave pipeline [13,17]. The faults at individual stages can be represented with the help of Jelinski-Moranda model as shown in Figure 5.

In the proposed method the data propagation is monitored at every stage with special control circuitry to enhance the accuracy. Even when the clock frequency is not synchronized the pipeline stages will control the previous data through gate logic as shown in Figure 1. It is observed that data pulses are propagated accurately when compared with other existing methods. An accurate data waves are observed at different timing rates as shown in Figure 6.

To maintain higher performance of the pipeline predefined accuracy levels and different failure rates are defined. For materializing a mathematic equation and for evaluating reliability following assumptions are made.

The system contains N homogeneous stages and failure density depends on number stages and it is exponentially distributed for i=*n.* In the present system, λ_c is the failure rate due to the clock skew, λ_p is the failure rate due to the delay difference of logic path between stages, λ_g is the failure rate due to the gate logic control. Gate logic control is used to control the data wave propagation to the next stages. R(t) is the reliability of the pipeline at individual stage at predefined sample size. The reliability function is evaluated as,

$$R(t) = \sum_{i=0}^{n} P_i(t) \qquad \text{---------- (6)}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The probability of failure rate distribution due to λ_c,λ_p, and λ_g, is P(t),

$$P_i(t) = \left(\lambda_i\left(N-i+1\right)+\lambda_p+\lambda_g+\lambda_c\right)e^{\left(-\lambda_i\left((N-i+1)+\lambda_p+\lambda_g+\lambda_c\right)t_i\right)} \quad \text{---------- (7)}$$

$$where, i = 1,2,3,...N$$

The reliability between first and second stage is R_{p12} and reliability of data propagation between second and third stage will be R_{p23} and so on, and R_{p12}, R_{p23} are in series.

Then the total reliability is given by=

$$R_{p12} = \left[1-\prod_{i=1}^{2}(1-R_i)\right]$$
$$R_{p23} = \left[1-\prod_{i=2}^{3}(1-R_i)\right] \quad \text{---------- (8)}$$

$$R(t) = R_{p12}R_{p23} \quad \text{---------- (9)}$$

For n stage pipeline the overall Reliability is,

$$R(t) = \sum_{i=1}^{n} R_{pi(i+1)} \quad \text{---------- (10)}$$

3. Result Analysis

The circuits for respective pipelines are constructed in Proteus and Electronic Work Bench. The reliability of the circuit design is tested, modified and analysed in Proteus.

The results are obtained and analysed with Electronic Work Bench to analyse data throughput. The results are obtained as shown in Table 1 at different frequencies.

A four stage pipeline is constructed and analysed for n=4. The reliability model is designed and mathematical formulas are evaluated in section 2. An accurate data waves are observed and so system design is reliable.

4. Conclusion

The parameters are observed on a four stage pipeline and assumed number of tasks is equal to 4. The performance is analysed at different data rates starting from 5Hz to 1GHz. In this paper readings at higher data rates are represented in the table and graph. It is observed that the new method showing optimistic results in timing delay, throughput and average delay when compared with other three methods. But in case of wave pipeline it is observed better results than traditional and Mesynchronous pipeline, with fewer logic gates. And still the new method need to be observe by cascading higher stages. The reliability of the proposed method is found high by considering failure rate λ_g, and other failure rates.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

5. Reference

[1] Gay CT. Timing constraints for wave pipelinedsystems. IEEE transactions on Computer aided design of integrated circuits. 1994 Aug, 13 (8), pp. 987-1004.
[2] Kumar NS, Reddy DV. Effect of Interrupt Logic on Delay Balancing Circuit. International Journal of Computer Applications. Published by Foundation of Computer Science New York, USA. 2011 Aug, 27(4), pp. 26-30
[3] TOKHEIM RL. Schaum's Outline of Theory and Problems of Computer Architecture Copyright © The McGraw-Hill Companies Inc. Indian Special Edition. 2009, pp. 1-17.
[4] Chang CH, Davidson ES. Delay balancing using Latches. 2000, pp. 1-9.
[5] Hwang K, Briggs FA. Computer Architecture and parallel processing. McGraw-hill International Editions, Computer Science Series, 1st edition. 1990.
[6] Kumar NS, Reddy DVRK, Sridevi R, Sridevi V. Studies and a Method to Minimize and Control the Jitter in Optical Based Communication System. IJACSA. 2013, 4 (4), pp. 45-51.
[7] Huang HY, Cheng KH, Wang JS, Chu YH, Wu CY. LOW-voltage Low-Power CMOS True Single-Phase Clocking scheme with Locally Asynchronous Logic Circuits. IEEE Xplore. 1995, 3, pp. 1572-1575.
[8] Shrivishwanath D, Raajan NR, Gopinath L. Flip Flop and Double Substitution Based Textual Encryption. Indian Journal of Science and Technology. 2014 Apr, 7 (4), pp. 52-55.
[9] Vignesh B. Pipelined Quadratic Equation Based Novel Multiplication Method for Cryptographic Applications. Indian Journal of Science and Technology. 2014 Apr, 7 (4), pp. 34-39.
[10] Kumar NS, Reddy SVRK. A New Method to Enhance Performance of Digital Frequency Measurement and Minimize the Clock Skew. 2011, 11 (10), pp. 2421-2425.
[11] Elavarasi R, Kumar PKS. An FPGA Based Regenerative Braking System of Electric Vehicle Driven by BLDC Motor. Indian Journal of Science and Technology. 2014 Nov, 7 (7), pp. 1-5.
[12] Gray CT. Timing constraints for wave pipelined systems. IEEE transactions on Computer aided design of integrated circuits. 1994 Aug, 13 (8), pp. 984-1004.
[13] Burleson WP, Ciesielski M, Klass F. Wave pipelining: a tutorial and research survey. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 1998 Sep, 6 (3), pp. 464 -474.
[14] Tatapudi SB. A Mesychronous high performance digital systems. 2006 May, 53 (5), pp. 1-19.
[15] Kumar NS, Reddy DVRK. Measurement system for wide range frequency at nonlinear devices. FCICS Published by Taylor and Francis group. 2015, pp. 143-146.
[16] Joe H, Reid N. On the software reliability models of Jelinski-Moranda and Littlewood. IEEE Transactions on Reliability. R. 1985, 34 (3), pp. 216-218.
[17] Rao YN, Raju SV. Enhancing Data Fetching Rates with Parallel Pipeline. International Journal of Computer Applications. 2014, 85, pp. 31-34.

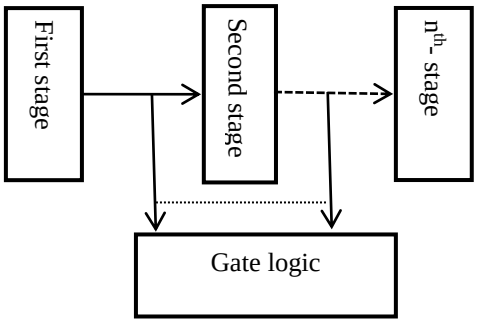


Figure 1. Proposed Method with new clock scheme

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

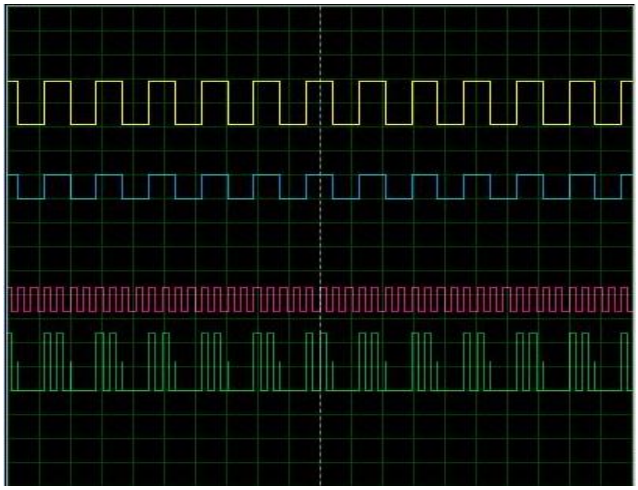


Figure 2. Presence of Failure in traditional pipeline

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

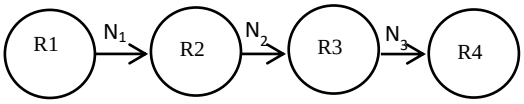


Figure 3. Fault assumptions of Jelinski-Moranda model for traditional pipeline

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

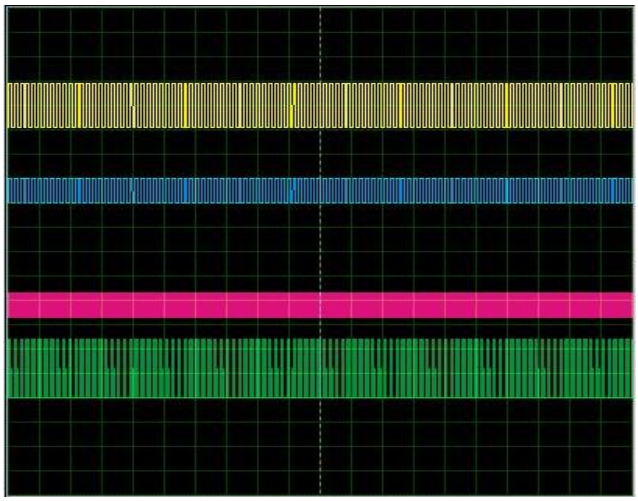


Figure 4. Data Propagation through Wave pipeline Clock scheme

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

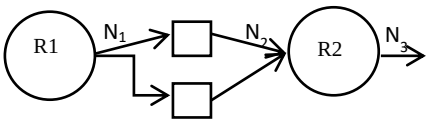


Figure 5. Fault assumptions of Jelinski-Moranda model for Wave Pipeline

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

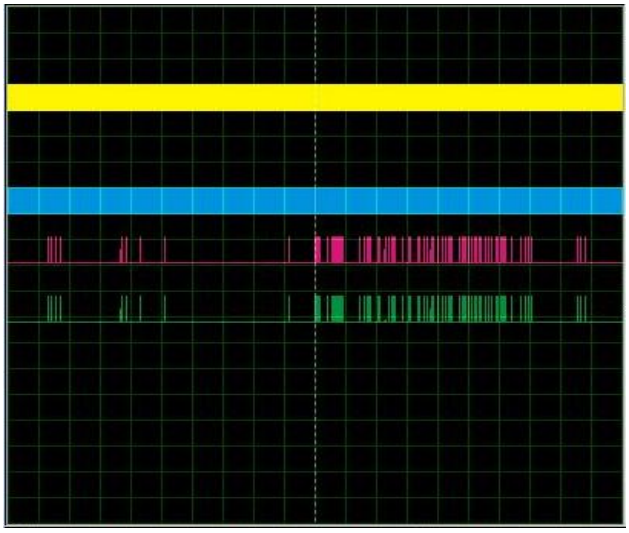


Figure 6. Data Propagation through New method

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1. Results of different four stage pipelines

Data Frequency in MHz	Traditional pipeline			Wave Pipeline			Mesochronous Pipeline			Proposed Pipeline		
	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τavg) in µsecs	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τ_{avg}) in µsecs	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τ_{avg}) in µsecs	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τ_{avg}) in µsecs
1000	1.7	2.3	0.425	1.8	2.2	0.45	1.7	2.3	0.425	1.4	2.86	0.35
500	1.78	2.25	0.445	1.78	2.25	0.445	1.67	2.4	0.418	1.44	2.77	0.36
100	1.89	2.12	0.473	1.89	2.12	0.473	1.78	2.25	0.445	1.62	2.47	0.405

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Efficient Voice Based Person Identification System for Secured Communication

Dr.Shakila Basheer

Lecturer in Information system Department

King Khalid university

ABHA

Saudi arabia

Abstract

Secured Communication is essential due to scalability due to increase number of devices and drastically growing number of people involved in communication. In this paper a voice comparison based communication authentication mechanism is used for providing secured communication. This voice based authentication is used in two different applications like people communication and data retrieval. Before going to speak with people in online their information and their voice is compared and verified from the database and permission will be granted. Similarly according to the voice they can retrieve the data from the data base, where it provides data integrity. Both applications comprise a number of stages such as: (i) Voice, Voice to Text input, (II). Voice Comparison and Pattern Matching. Finally (III). Permission Granted and Data Retrieval (DR) as the output. In order to improve the accuracy and relevancy the proposed data retrieval system, it uses an indexing method called Bag of Words (BOW). BOW is like an index-table which can be referred to store, compare and retrieve the information speedily and accurately. Index-table utilization in DRS improves the accuracy with minimized computational complexity. The proposed DRS is simulated in DOTNET software and the results are compared with the existing system results in order to evaluate the performance.

Keywords: Information Retrieval System, Data Mining, Bag of Words, Data Base Maintenance.

Introduction

In general IR is an activity is used by a few people for library management, paralegals and the digital library searching system. The world is growing with lots of changes were more than million number of people are using IR in everyday life like email, web searching. After sometime the IR system is used for information access and traditional searching in databases such as, searching an order, searching a product, searching a document from a digital library and so on. It is well known that the IR retrieves data from unstructured databases. The term “unstructured data” means the data is not clear, semantically overt and the format of the data is undefined. Simply can say that it is opposite to structured data (example: DBMS, RDBMS), but in real-time there is no data are not truly unstructured.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Searching information, images, documents and files are created based on the visual appearance and the properties of the data, document and images. Information retrieval is a challenging problem where it has been received a considerable attention from most of the researchers in various fields of image processing, data mining, information retrieval and computer vision and multimedia systems. The growth of web technology brings a drastic increase in data usage published in the recent decades, which has been a great challenge to develop efficient information retrieval systems to help all the users in IR systems. Traditional IR models such as: vector space model [16], classical probabilistic IR models [15] and language modeling approaches [13] are used for query based document retrieval and works independently. Web search engines are used for entity based retrieval [14, 12] used for commercial purpose. An entity based web document retrieval [9-11] are used in the earlier research works to provide a better semantic based document searching. Searching, information retrieval, content based information retrieval systems are still getting urgent demand in the web applications [17,18]. The retrieval system concentrates on features as important for information extraction. Most of the paper follows the feature based IR on content based image retrieval systems [19-21]. Some of the IR systems used to transform in order to decompose and represent various resolutions, various sizes and various amounts of information [22-23]. Wavelet transform have been successfully applied to image Denoising [24], image compression [25] and texture analysis [26]. In [27] the authors propose a new CBIR system using color and texture features. In this paper texture features are extracted Euclidean distance measure to obtain the similarity measurement between a query text and text in the database. In [28] wavelet basis was used to characterize each query image and also to maximize the retrieval performance in a training data set. To make DRS is more efficient, DRS is not constructed based on all the entities. It is query independent. For each voice query the index is selected and then the related data are selected from different location. One the index is matched, and then DRS decides the location of the data and the entities of the index-data from the database. In this paper the information retrieval system is developed using index searching and pattern matching methodologies. To do index searching BOW is used. The contribution of the proposed DRS work is:

- Voice (input)
- Creating BOW
- Voice Matching and Pattern Matching
- Communication Permission granted and Data Retrieval (output)

Proposed Model

The proposed model clearly says about the entire functionality of the proposed DRS and it is shown in Figure-1. Any physically challenged people one who are not able to operate the keyboard can use this application. In this paper, it is assumed that the application is developed for online shopping. The user can say about the product in mic then the voice is converted into text. The converted text is taken as a keyword for pattern matching in the product database. During the pattern matching keyword is verified with the BOW in order to check the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

product availability. If the keyword is available in BOW then the other relevant information about the product is taken from the database, converted into voice, and play back to the user. It is an advanced application can be used in handheld devices also.

In user communication, initially the numbers of users are registered with their voice. The voice is the keyword for comparison, whereas before coming to communicate in online both end user has to be verified by the voice. If the present users' (ready for communication) voice is matched with the stored DB voice then they are permitted for communication and they can proceed. The this functionality is depicted in Figure-1.

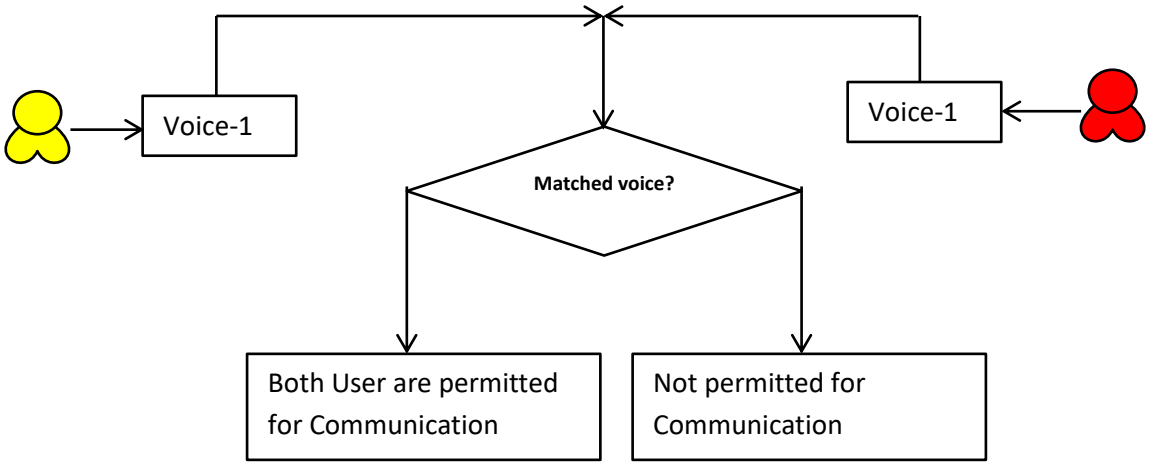


Figure-1: Application-1 [Secured User Communication]

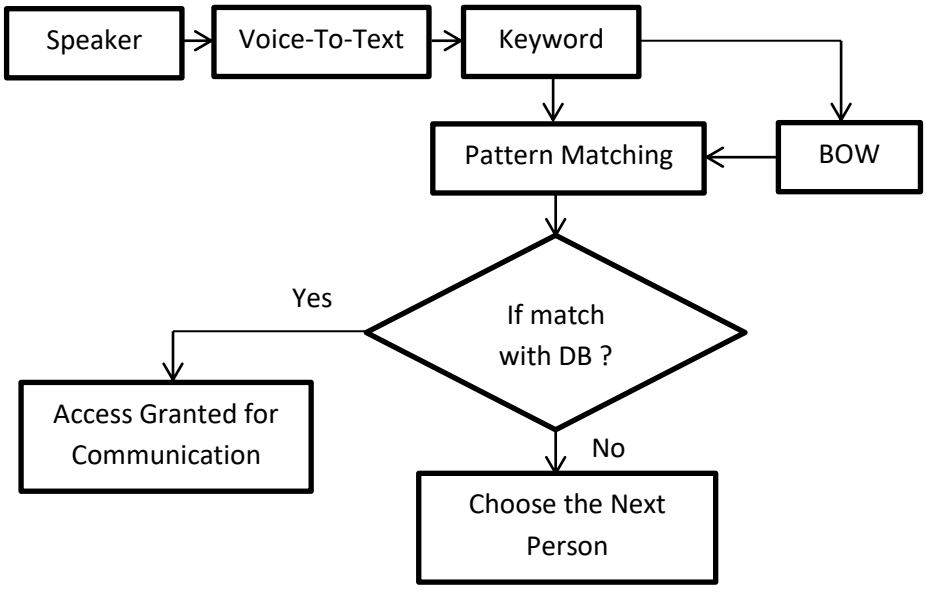


Figure-2: Application-2 [Secured Data Retrieval]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Bag-of-Words

One of the most common methodologies to obtain the entire data is by visual words and it can be applied as text indexing and retrieval scheme. The index is created from any one of the feature of the data stored in the DB before persisting newly in the DB. It can be called as Bag-of Words or bag of feature model. Some of the static terms are taken from the data and it is maintained as a catalog (BOW). This catalog is compared with the database data for retrieving the specific data matched with the catalog. The data retrieval using keywords can predict maximum relevant based data and it satisfies the customer.

In this proposed DRS, whenever a new product detail is entered into the database, any one of the data feature is added as index word into the BOW. It needs not be a numeric or character data type and it can arrange the entire BOW automatic while inserting a new index. This automatically arranging of index words helps to compare and retrieve the relevant data speedily and accurately without computational complexity. For example: when a new data d_n is inserted into the database D, one of the feature from the feature set $f_i = \{f_1, f_2, \ldots, f_n\}$ is stored into BOW.

Field-1	**Field-2**	**…….**	**Field-i**	**……**	**Field-n**
D_1f-1	D_1f-12	…….	D_1f-i	……	D_1f-n
:	:	:	:	:	:
:	:	:	:	:	:
D_nf-1	D_nf-2	…….	D_nf-i	…….	D_nf-n

BOW
D_1f-i
:
:
D_nf-i

Figure-3: BOW Creation

Each field of the data is considered as separate features and any one of the field is stored into BOW. In an image retrieval system BOW is created automatically using LABELME tool. But in case of alphanumeric data the feature has chosen as keyword manually by the developer according to the convenient. Figure-3 shows the way of BOW creation and it can be used to compare the product availability in the database.

The data classification and retrieval is based on the BOW index, where BOW is the structured features taken from all the trained data inserted in the database. The word stored in the BOW belongs to the same class and it is behaving like a codebook used to cluster and classify the entire dataset. The words of all dictionaries represent frequent structures of all form types. Each word type is represented by a feature vector. The structural features of a form $\mathbf{s_j}$ are

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

calculated and are assigned to the cluster center $\mathbf{w_i}$ (word) with the smallest (Euclidian) distance mini $\|\mathbf{s_j}\text{-}\mathbf{W_i}\|$. This distance is used to fetch the matching BOW for the voice into text (keyword).

Voice –To-Text

A portion of the DRS system is programmed to recognize the speech (voice), and convert into text using speech synthesization mechanism available in the system library. The inbuilt speech recognition engine is instantiated initially, then the defined grammar is loaded in order to recognize the phrases. Adding grammar is used to identify the grammar-name. Each time the grammar is loaded dynamically in order to update the new BOW inserted. This updating can be obtained by the recognizer update method. In this paper the DRS listens to the user whether any speech data is entered into the system. The speech recognition engine is already loaded with the predefined trained text in the background. Each time speech made one line of text is displayed at a time in the system. The main advantage of this system is it will wait for a small interval in order to avoid congestion and proceed with the next BOW. If the speech is understandable by the speech engine then it keeps idle and waits for the next speech and it won't create any software breakup.

The speech to text is an application where it does translate words into text as much as possible due to various countries' accent variation. Other than the DRS, this voice to text conversion is used in healthcare, traffic systems, military, telephony and education systems. It is mainly focused for people with dis-abilities. This paper follows a fuzzy logic based Speech Recognition of Linguistic Content method [1]. In this method a word in a language, speaks in different accents, different speeds of pronunciation and with different emphasis. For example, the word "vector" of the English language will be spoken by an American as "vektor", with curtness at the 'c' and at the 't', while a Britisher will speak it as "vectorr", with emphasis on the 'c' and a slight repetition on the 'r'. Similarly, a Russian will speak this word as "vecthor", with softness on the 't'. However, the word remains the same, that is, "vector", with slight variations with respect to different accents, speeds of pronunciation and emphasis.

Thus, a single word can be represented as a fuzzy set. However, a word is too specific so as to fit into a generic model of speech recognition. To have a more general model, the fuzzification of phonemes is more appropriate. This model is therefore applied to spoken sentences. One fuzzy set is based on accents, the second on the speeds of pronunciation and the third on emphasis. The use of this method will be especially for speech-to-text conversion, by filtering out the unnecessary paralinguistic information from the spoken sentences.

Pattern Matching

In this paper the main idea is to search from right to left in the pattern. With this scheme, searching is faster than average. In order to do this the Boyer-Moore (BM) algorithm positions the pattern over the leftmost characters in the text and attempts to match it from right to left. If no mismatch occurs, then the pattern has been found. Otherwise, the algorithm computes a shift;

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

that is, an amount by which the pattern is moved to the right before a new matching attempt is undertaken. The shift can be computed using two heuristics: the match heuristic and the occurrence heuristic. The *match* heuristic is obtained by noting that when the pattern is moved to the right, it must

1. Match ***all*** the characters previously matched, and

2. To bring a different character to the position in the text that caused the mismatch.

The last condition is mentioned in the Boyer-Moore paper [3], but was introduced into the algorithm by Knuth et al. [2]. Following the later reference, we call the original shift table $\boldsymbol{dd}$, and the improved version $\widehat{\boldsymbol{dd}}$. The formal definitions are

$$\widehat{\boldsymbol{dd}}[\boldsymbol{j}] = \mathbf{min}\{\boldsymbol{s} + \boldsymbol{m} - \boldsymbol{j} | \boldsymbol{s} \geq \mathbf{1}\ \boldsymbol{and}\ ((\boldsymbol{s} \geq \boldsymbol{i}\ \boldsymbol{or}\ \boldsymbol{pattern}[\boldsymbol{i} - \boldsymbol{s}] = \boldsymbol{pattern}[\boldsymbol{i}])\ \ \boldsymbol{for}\ \boldsymbol{j} < \boldsymbol{i} \leq \boldsymbol{m})\}$$

$$\boldsymbol{for}\ \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{m}; \boldsymbol{and}$$

$$\widehat{\boldsymbol{dd}}[\boldsymbol{j}] = \mathbf{min}\{\boldsymbol{s} + \boldsymbol{m} - \boldsymbol{j} | \boldsymbol{s} \geq \mathbf{1}\ \boldsymbol{and}\ ((\boldsymbol{s} \geq \boldsymbol{j}\ \boldsymbol{or}\ \boldsymbol{pattern}[\boldsymbol{j} - \boldsymbol{s}] \neq \boldsymbol{pattern}[\boldsymbol{j}])\ \boldsymbol{and}\ ((\boldsymbol{s} \geq \boldsymbol{i}\ \boldsymbol{or}\ \boldsymbol{pattern}[\boldsymbol{i} - \boldsymbol{s}] = \boldsymbol{pattern}[\boldsymbol{i}])\ \ \boldsymbol{for}\ \boldsymbol{j} < \boldsymbol{i} \leq \boldsymbol{m})\}$$

The $\widehat{\boldsymbol{dd}}$ table for the pattern $\boldsymbol{abracadabra}$ is

dd	**a**	**b**	**r**	**a**	**c**	**a**	**d**	**a**	**b**	**r**	**a**
$\widehat{\boldsymbol{dd}}$ **[j]**	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**13**	**12**	**4**	**1**

The occurrence heuristic is obtained by noting that we must align the position in the text that caused the mismatch with the first character of the pattern that matches it. Formally calling this table *d*, we have

***d*[*x*] = min{*s*|*s* = *m* or (0 ≤ *s* < m *and* pattern *[*m - s*]* = x*)}**

for every symbol ***x*** in the alphabet. This methodology is used to compare the voice converted text with BOW and with the database. If the pattern matches the database, then the voice based reply is produced to the physically challenged people. The voice is produced by converting the relevant record information obtained from the database and convert into voice.

Text-To-Voice

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Text-to-speech synthesis takes place in several steps. The TTS systems get a text as input, which it first must analyze and then transform into a phonetic description. Then in a further step it generates the prosody. From the information now available, it can produce a speech signal. The structure of the text-to-speech synthesizer can be broken down into major modules:

Natural Language Processing (NLP) module: It produces a phonetic transcription of the text read, together with prosody.

- Digital Signal Processing (DSP) module: It transforms the symbolic information it receives from NLP into audible and intelligible speech. The major operations of the NLP module are as follows:

- Text Analysis: First the text is segmented into tokens. The token-to-word conversion creates the orthographic form of the token. For the token "Mr" the orthographic form "Mister" is formed by expansion, the token "12" gets the orthographic form "twelve" and "1997" is transformed to "nineteen ninety seven".

- Application of Pronunciation Rules: After the text analysis has been completed, pronunciation rules can be applied. Letters cannot be transformed 1:1 into phonemes because the correspondence is not always parallel. In certain environments, a single letter can correspond to either no phoneme (for example, "h" in "caught") or several phoneme ("m" in "Maximum"). In addition, several letters can correspond to a single phoneme ("ch" in "rich"). There are two strategies to determine pronunciation:

In dictionary-based solution with morphological components, as many morphemes (words) as possible are stored in a dictionary. Full forms are generated by means of inflection, derivation and composition rules. Alternatively, a full form dictionary is used in which all possible word forms are stored. Pronunciation rules determine the pronunciation of words not found in the dictionary.

In a rule based solution, pronunciation rules are generated from the phonological knowledge of dictionaries. Only words whose pronunciation is a complete exception are included in the dictionary. The two applications differ significantly in the size of their dictionaries. The dictionary-based solution is many times larger than the rules-based solution's dictionary of exception. However, dictionary-based solutions can be more exact than rule-based solution if they have a large enough phonetic dictionary available.

Whenever a voice input into DRS it is taken as the query for searching the relevant product from the database. Query enriches expansion is a general strategy used in text retrieval, which is directly adapted to the BOW model in all kinds of data retrieval. In this project the query expansion is simply taken as index searching with BOW and pattern matching with the database. There are various query methods are available like Transitive Closure Expansion (TCE) [4], and Additive Query Expansion (AQE) [5]. In this paper the TCE is used for query

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

processing system. Initially the query word (voice to text) is compared with the index where each visual word has an index indicating that the entire data is available in the database or not. This paper doesn't calculate the score value defining the similarity [6], since the keyword is unique. Using the above text to speech conversion the voice reply is generated and play with the user. The entire functionality of the proposed DRS is given in the form of algorithms, it can be coded in any computer programming language and the efficiency can be evaluated.

Algorithm_DRS (string product)

{

Input: voice, product data, initial BOW;

Output: voice

Description:

1. user speech in mic
2. Voice is converted into text
3. Apply a pattern matching algorithm
4. Search text into BOW
5. If (text exists in BOW) then search in DB
6. Voice (" product details"); // all the fields from the matched field is converted into voice
7. Else
8. Voice ("product not available");
9. End
10. If any product insertion then
11. field-i insert into BOW

}

Experimental Setup

The functionality of the proposed DRS is programmed in DOTNET 2010 software and the results are produced. There are 25 systems are installed in a laboratory in order to evaluate the system performance. In all, the system DOTNET software and the IRS module is installed. The proposed DRS are programmed, experimented in DOTNET software and the results are given below to analyze the performance. One among the systems is assumed as the server and the database is installed. The database is a lexical dictionary which consists of a collection of data in the form of rows. Each row consists of various numbers of columns which is not having appeared like a table. Another system is assumed as a middleware, having BOW table, which consists of a set of all inserted index keywords. Whenever a voice input entry to the system it

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

refers the BOW first and then comes to the database server, which reduces the computational complexity.

In order to experiment the proposed DRS, a product dataset is taken from [8] and experimented. 100 different products are stored in the database. It is assumed that the most of the product names are known by the user and it is online shopping. Some of the product name with some more relevant information about the product is shown in Table-1. Product code, product name are the two main features mostly used for searching the product information speedily in the entire database. Instead of concentrating all modules of online shopping, it is simply coming to know the product availability and product price with other relevant information about the product. The database consists of 15 fields in the table were on our paper only 5 fields are taken as important information to verify the DRS performance. In common product-code is used as searching indexes, but here due to voice mining, product name is used as searching indexes.

Table-1: Product Information

productCode	productName	productLine	quantityInStock	buyPrice	productDescription
S10_1678	1969 Harley Davidson Ultimate Chopper	Motorcycles	7933	48.81	This product is good and u can get world service
S10_1949	Alpine Renault 1300	Classic Cars	7305	98.58	Turnable front wheels; steering function; detailed interior; detailed engine; opening hood; opening trunk; opening doors; and detailed chassis
S10_2016	1996 Moto Guzzi 1100i	Motorcycles	6625	68.99	detailed engine, working steering, working suspension, two leather seats, luggage rack, dual exhaust pipes, small saddle bag located on handle bars, two-tone paint with chrome accents, superior die-cast detail , rotating wheels , working kick stand

There are 100 data is stored in the table where during searching computational time is spent only 100 comparisons and data fetching. For **an N number** of comparisons the computation time taken is **2N+2.** The following figures show that the efficiency of the proposed DRS in terms of accuracy, timeliness and response generations. In order to evaluate the performance, the number of data used in the database table is changed and verified. The number of data is changed from 100 to 1000 and the performance is compared.

In this paper the user provides their input as voice through multimedia input device. The voice is recorded and recognized by the speech engine installed in the system and it is converted into text. The voice recognition is a big process if the Voice-accent is understood by the speech

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

engine then it converts the voice into text. In this process, the number of voices is recognized accurately for the voice input given into the DRS. In order to evaluate the voice recognition accuracy by the DRS, the number of voice input is increased and the recognition rate is calculated. The number of voice input may be changed from 25 to 250. Each round of experiments the number of voice input is increased by 25. Out of the input voice, the number of voices recognized by the DRS system is calculated and shown in Figure-3. Still Google-Voice play is also finding difficulties in terms of voice recognition. In the proposed DRS system the recognition rate is better and it is increased according to the number of voice input increases. The recognition rate is proportionally increased, according to the number of voice inputs getting increased. After successful recognition, the voice is converted into text (it is taken as a keyword) for comparison with the BOW. If the keyword matched with the BOW index, then directly compared with the database in order to process the pattern matching.

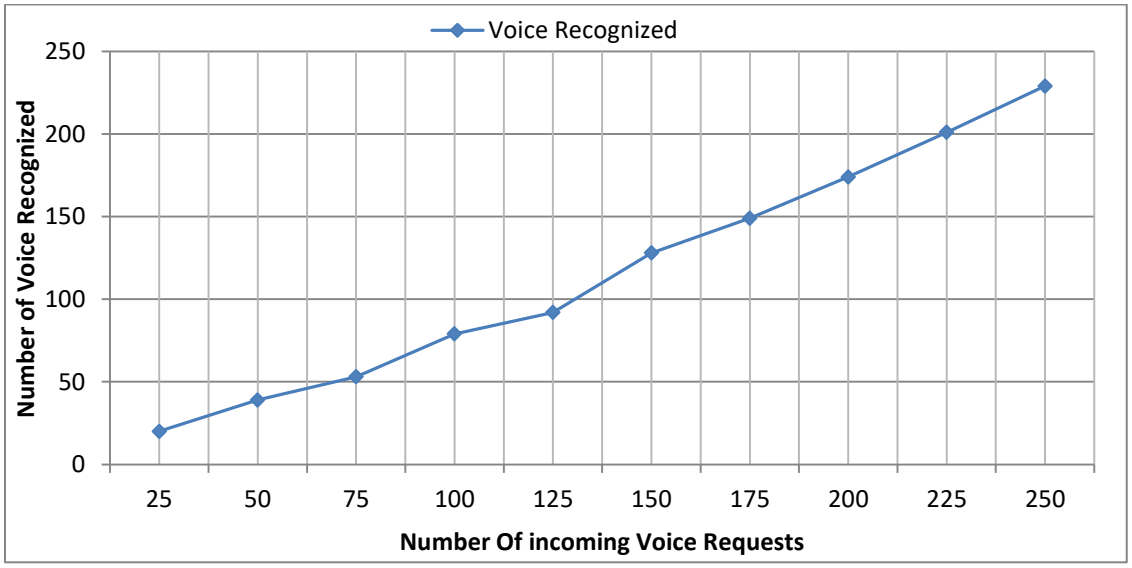


Figure-4: Number of Voice Inputs Recognized Vs. Number of Voice Inputs

If the pattern matched, then the relevant record data are fetched from the data row then converted into voice again. This text-to-voice conversion is played to the user who passed the voice input. According to the number of voice input processed, the number of voice reply is calculated and the quality of the DRS is verified. The number of voice reply against the number of voices is shown in Figure-5. Figure-5 says that the voice reply is increased according to the number of input voice. It is clear that after index matching the reply can be generated according to the pattern availability. The reply may be about the product or it is a message saying that particular product is not available and since there should be a compulsory voice reply for each voice input if it matched with the index. The execution process is preceded when the index is matched, else it is dropped executing the next process. Hence the proposed DRS reduces the computational complexity.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Also Figure-5 shows that the number of voice reply is merely equal to the number voice input given into the proposed DRS. It cannot be concluded that the pattern matching will be performed if the keyword matched with the BOW index due to the product may not be available. The pattern matching algorithms used in this paper find the distance between the possible patterns obtained from the DB with the input pattern. If the distance is merely equal to zero, then the pattern is matched, else it is not matched. According to the pattern matching algorithm, the accuracy is calculated and shown in Figure-6. The percentage of pattern matching is merely equal to the percentage of index matching. From this figure, it is clear that the number of pattern matching is lesser than the number of index matching. After the index matching successful the appropriate pattern may not available in the database and it affects the pattern matching accuracy. It cannot be concluded that the accuracy of the DRS is less. In this paper the accuracy of the entire IR system can be taken as the average of both index matching and pattern matching.

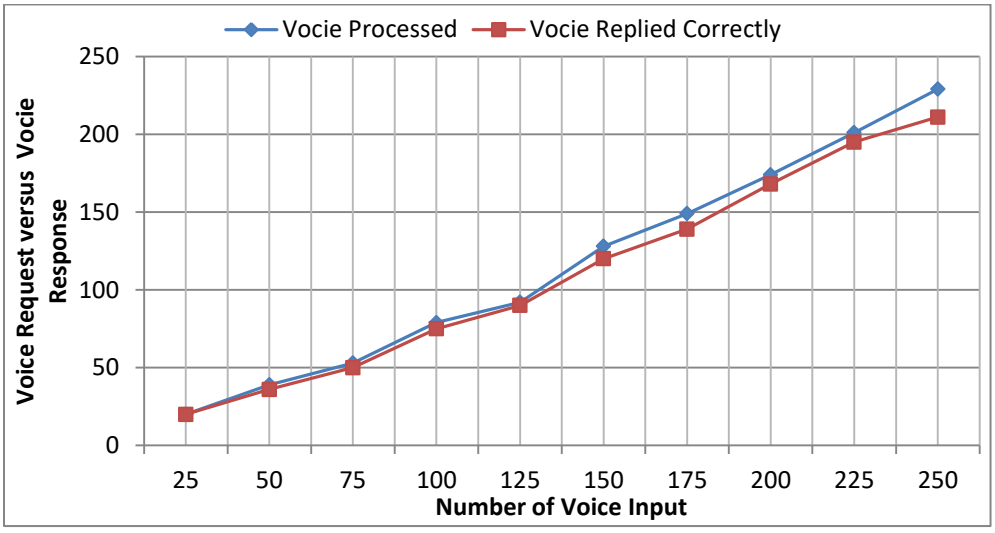


Figure-5: Number of Voice Input vs. Number of Voice Output

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

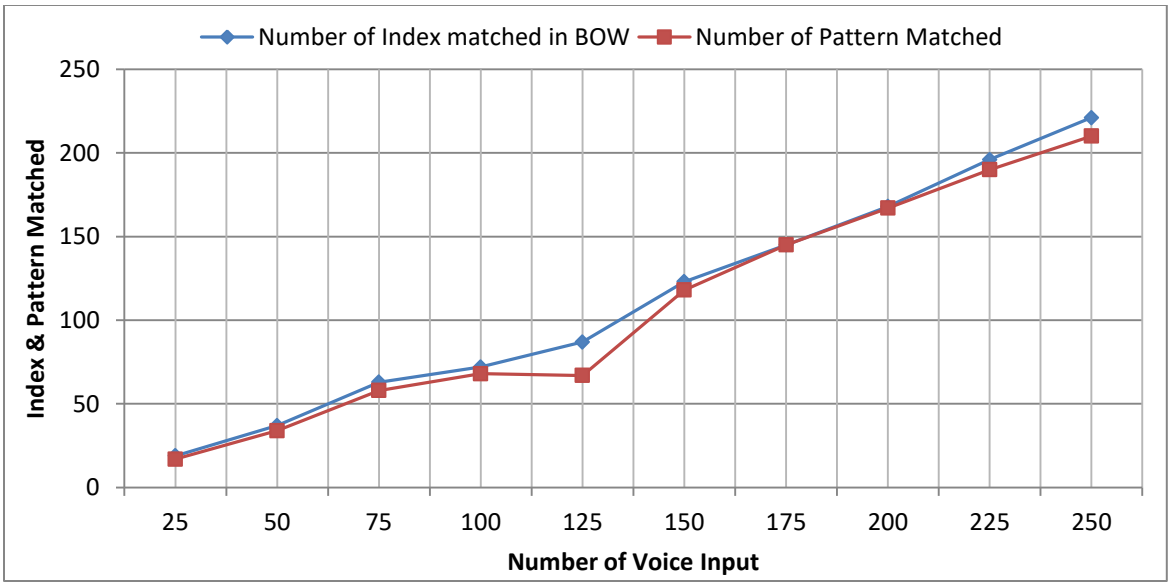


Figure-6: Voice Input Matched with Index and Matched With Pattern Comparison

The computational complexity refers the number of statements in the program to be executed in the compiler and the time taken to compile. The number of statements in the program decides the compilation time and the compilation time taken by the proposed DRS is shown in Figure-7. The figure shows that the computational time is less and it increases, according to the number of inputs increased. It means that for 100 numbers of data it takes only 4 seconds to make the entire process of DRS.

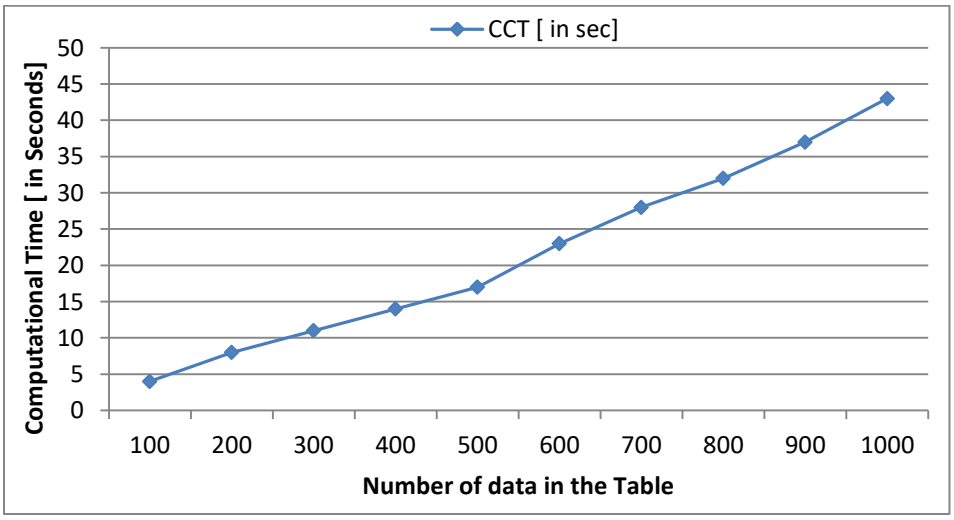


Figure-7: Computational Time In Terms of Data Size

Also the efficiency can be calculated according to the number of response generation against number of input queries. The number of query response against the number of input queries is shown in Figure-9. DRS proved that the number of pattern matching is not depending on the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

number of index matching completely. It depends on the index matching and the data availability. This figure shows the number of voice reply (response) provided to the user against the query input. The voice reply is gradually increased according to the number voice query applied. The accent and the data availability determine the accuracy of the pattern matching and voice reply accuracy.

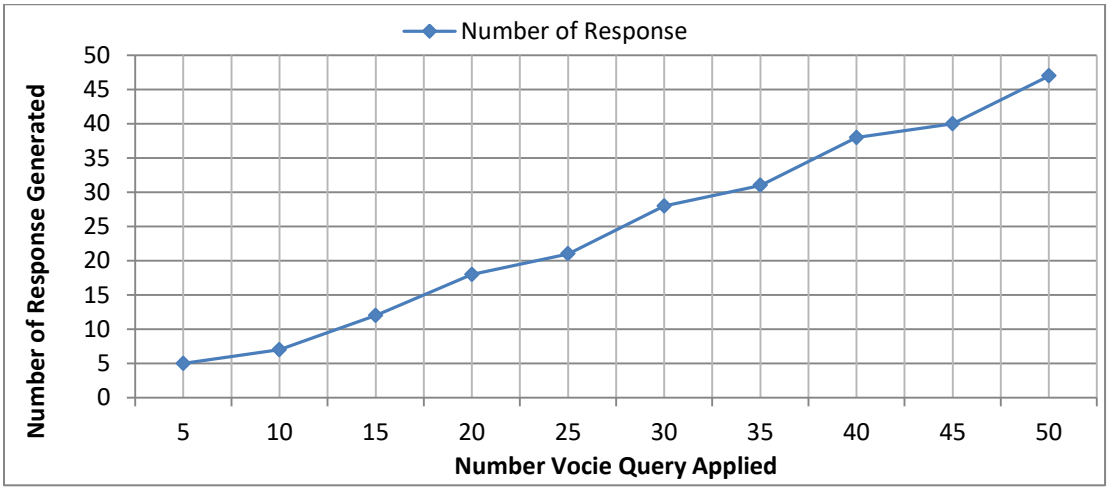


Figure-8: Number Query vs. Number of Response Generated

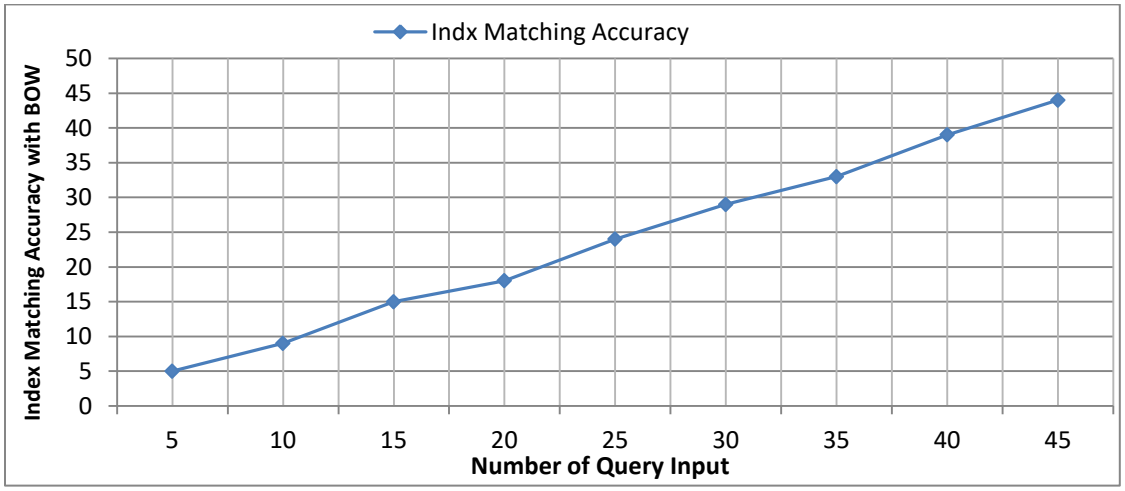


Figure-9: Number of Query vs. Index Matching Accuracy

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

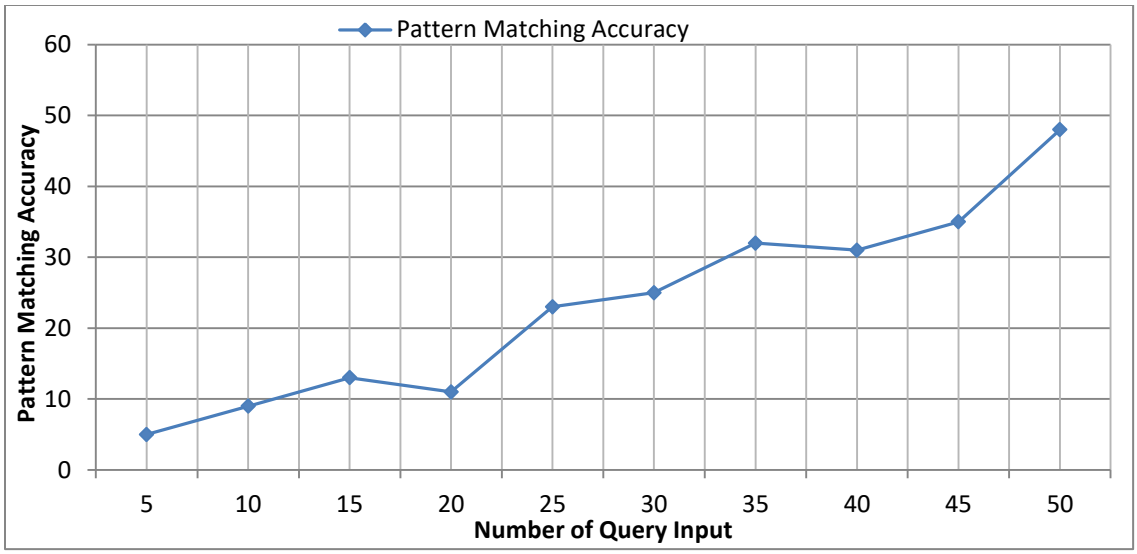


Figure-10: Number of Query vs. Pattern Matching Accuracy

In this paper the number of indices matched and the number of patterns matched is calculated and shown in Figure-9 and Figure-10 respectively. The number of query index matching is proportionally increased, according to the number of query data and accent. The number of pattern matching is up and down in scale due to match pattern and the data available on the DS. In order to evaluate the performance the proposed DRS results are compared with the existing approach.

Performance Analysis

The performance of the DRS is evaluated by comparing the mining accuracy and time complexity with the existing approaches [8]. The proposed DRS and the existing IR system are using the data-dictionary at the back end. The data dictionary size is 100, 120 and 140 in terms of number of words. Figure-10 shows the mining accuracy comparison between proposed DRS and the existing IR [8] system. It is clear that the mining accuracy obtained by the proposed DRS is more than the existing IR. To verify the accuracy and comparability the size of the data dictionary is changed gradually and experimented. In each time of the experiment the mining accuracy is also gradually increased in proposed DRS and it is greater than the existing IR accuracy. Time taken to process the query and response generation and for pattern matching is computed for the proposed DRS and compared with the existing IR system. The time taken by the proposed DRS is lesser than the existing approach time. The experiment is repeated for all the dictionary size 100, 120 and 140, and the time calculated. The calculated time includes the voice processing, BOW index matching and pattern matching time. The complete processing time for one job in the proposed DRS is, time from query word is obtained from voice, compared with the BOW, if exists then it compare with the database. Time taken to process the information retrieved by the proposed and existing is shown in Figure-11. From this figure, it is clear that the time taken by the proposed approach is lesser than the existing approach.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

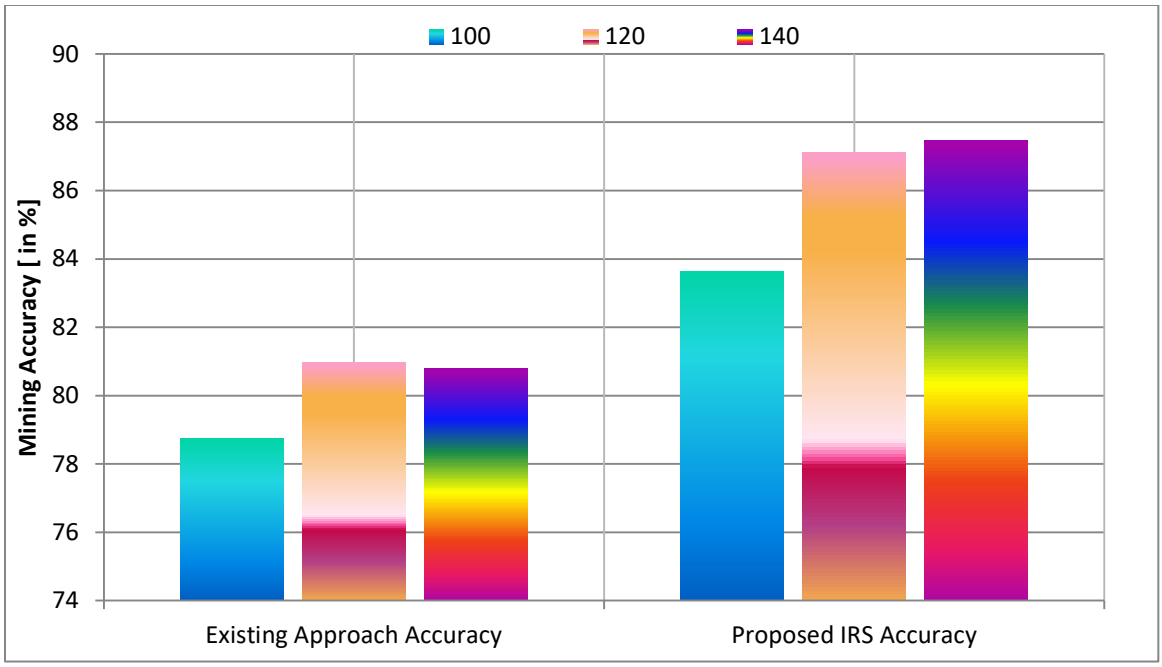


Figure-10: Data Mining Accuracy Comparison between Proposed DRS and Existing Approach

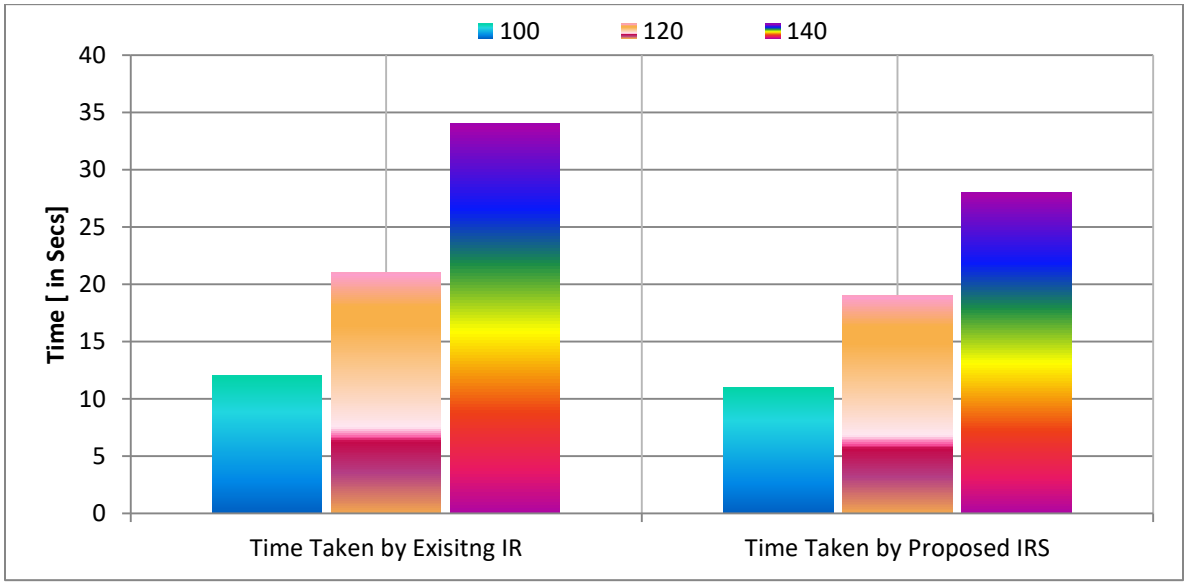


Figure-11: Time Comparison between Proposed DRS and Existing Approach

Run time Efficiency

The efficiency of the proposed DRS is calculated while applying DRS to provide online retrieval and voice reply for large set of database collection. Comparing with the traditional IR approaches, the overhead of proposed voice based DRS comprises four parts: (i). Converting voice to text; (ii). Matching query words in BOW; (iii). Pattern Matching with DB; (iv). Voice based Reply. The previous research off-the-shelf recognition toolkits could already handle the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

entity annotation on queries well with the high accuracy and low latency. By building BOW using the data features, the overhead of index matching and pattern matching process is reduced to do information retrieval. It reduces the time complexity and computational complexity and since this proposed DRS can be extended to large scale data collection, web applications and in wireless network based applications.

Conclusion

The main objective of this paper is to develop a data mining model for physically challenged people using voice. The proposed DRS method uses BOW model in order to retrieve the relevant information from the data. Comparing BOW reduces the computational complexity and searching time. In this paper the proposed DRS handle a smart way of information retrieval approach, which estimate the data availability by comparing the index in order to reduce the time and computational complexity. It can be applied for high – dimensional data entity space. This proposed DRS provides voice to text, text to voice and visual word comparison for improving the efficiency of the information retrieval system. From the results it is clear that this approach is efficient in term of reduced computation complexity, reduced time and it is a special kind of information retrieval system helps to social for physically challenge people like blind and no able to operate keyboard. This voice comparison based authentication can be utilized in various kinds of applications and it is proved.

Reference

[1]. Lakra, Sachin, et al. "Application of fuzzy mathematics to speech-to-text conversion by elimination of paralinguistic content." *arXiv preprint arXiv:1209.4535* (2012).

[2]. KNUTH, D., J. MORRIS, and V. PRATT. 1977. "Fast Pattern Matching in Strings." *SIAM J on Computing*, 6, 323-50.

[3]. BOYER, R., and S. MOORE. 1977. "A Fast String Searching Algorithm." *CACM*, 20, 762-72.

[4]. Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In ICCV, pages 1–8, 2007.

[5]. HHerv´e J´egou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. International Journal of Computer Vision, 87(3):316–336, 2010.

[6]. James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In CVPR, 2007.

[7]. http://www.mysqltutorial.org/mysql-sample-database.aspx.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[8]. Kleber, Florian, Markus Diem, and Robert Sablatnig, "Form classification and retrieval using bag of words with shape features of line structures"-*IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, 2013.

[9]. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In IJCAI, volume 7, pages 2670–2676, 2007.

[10]. M. J. Cafarella, J. Madhavan, and A. Halevy. Web-Scale Extraction of Structured Data. ACM SIGMOD Record, 37(4):55–61, 2009.

[11]. S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

[12]. T. Lin, P. Pantel, M. Gamon, A. Kannan, and A. Fuxman. Active Objects: Actions for Entity-Centric Search. In *WWW*, pages 589–598, 2012.

[13]. J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *SIGIR*, pages 275–281, 1998.

[14]. J. Pound, P. Mika, and H. Zaragoza. Ad-hoc Object Retrieval in the Web of Data. In *WWW*, pages 771–780, 2010.

[15]. S. E. Robertson and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR*, pages 232–241, 1994.

[16]. G. Salton, A. Wong, and C.-S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[17] Kherfi, M.L., Ziou, D. and Bernardi, A. (2004) Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Computing Surveys*, **36**, 35-67. http://dx.doi.org/10.1145/1013208.1013210

[18] Datta, R., Joshi, D., Li, J. and Wang, J.Z. (2008) Image Retrieval: Ideas, Influences, and Trends of the NEW Age. *ACM Computing Surveys*, **40**, 1-60.

[19] Yang, M., Kpalma, K. and Ronsin, J. (2010) A Survey of Shape Feature Extraction Techniques. *Pattern Recognition*,1-38.

[20] Penatti Otavio, A.B., Valle, E. and Torres, R.da.S. (2012) Comparative Study of Global Color and Texture Descriptors for Web Image Retrieval. *Int. J. Via.Commun. Image R*, 359-380.

[21] Deselaers, T., Keysers, D. and Ney, H. (2008) Features for Image Retrieval: An Experimental Comparison. *Information Retrieval*, **11**, 77-107.

[22] Mallat, S.G. (1989) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674-693.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[23] Sarck, J.L., Murtagh, F.D. and Bijaoui, A. (1998) Image Processing and Data Analysis: The Multiscale Approach.

[24] Hill, P., Achim, A. and Bull, D. (2012) The Undecimated Dual Tree Complex Wavelet Transform and Its Application to Bivariate Image Denoising Using a Cauchy Model. 19*th IEEE International Conference on Image Processing* (*ICIP*), 1205-1208. http://dx.doi.org/10.1109/icip.2012.6467082.

[25] Kalra, M. and Ghosh, D. (2012) Image Compression Using Wavelet Based Compressed Sensing and Vector Quantization. *IEEE* 11*th International Conference on Signal Processing* (*ICSP*), **1**, 640-645.

[26] Kokareh, M., Biswas, P.K. and Chatterji, B.N. (2005) Texture Image Retrieval Using New Rotated Complex Wavelet Filters. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*: *Cybernetics*, **35**, 1168-1178.

[27] Balamurugan, V. and Anandha Kumar, P. (2008) An Integrated Color and Texture Feature Based Framework for Content Based Image Retrieval Using 2D Wavelet Transform. *IEEE International Conference on Computing, Communication and Networking*, 1-16. http://dx.doi.org/10.1109/icccnet.2008.4787734

[28] Quellec, G., Lamard, M., Cazuguel, G., Cochener, B. and Roux, C. (2012) Fast Wavelet-Based Image Characterization for Highly Adaptive Image Retrieval. *IEEE Transactions on Image Processing*, **21**, 1613-1623.

Comparative analysis of modern methods and algorithms of cryptographic protection of information

Saleh I. Alomar

AL-Ahliyya amman university Department of Engineering

Saleh A. khawatreh

AL-Ahliyya amman university Department of Engineering

***Abstract*—Information protection problems are topical at the present stage of development of information technologies. Protection of information stored in electronic form, is implemented by cryptographic methods. The article deals with modern symmetric and asymmetric encryption methods. It analyzes advantages and disadvantages of each type of encryption algorithms. Based on comparison results of algorithms, recommendations on the use of algorithms to solve specific problems are provided. The aim of the article is to analyze modern methods and encryption algorithms. When analyzing the strengths and weaknesses of cryptographic methods of protection it is necessary to make a choice of the method of protection on the basis of selected performance criteria, as well as assess the possibility of practical use of the considered cryptographic protection methods for different tasks.**

***Keyword*—**cryptoalgorithm, symmetric algorithm, an asymmetric algorithm, ciphertext.

I. INTRODUCTION

Cryptography, over the ages, has been an art practiced by many who have devised ad hoc techniques to meet some of the information security requirements. The last twenty years have been a period of transition as the discipline moved from an art to a science.

Cryptography is the study of mathematical techniques related to aspects of information security such as confidentiality, data integrity, entity authentication, and data origin authentication.[6]

The constant increase in the volume of confidential information, appearing of new methods and means of unauthorized access to the data leads to the development of information security industry. It is reflected in the creation of new methods and improvement of the existing ones and cryptographic protection algorithms.

The essence of this deficiency is that in the process of breaking any of the known cryptographic systems the cryptanalyst is able to identify the moment their work is successfully completed. This ability stems from the fact that during the cryptographic enciphering of the text the semantic content of the information being protected is, as a rule, transformed into semantically undefined set of alphabet symbols used.[6]

Cryptographic methods and algorithms for protection of information can be divided into:

-symmetric cryptosystem

-asymmetric cryptosystems

Each type of encryption algorithm has its own specific implementation features, advantages and disadvantages that must be taken in dealing with specific problems.

-Symmetric encryption is the oldest and best-known technique. A secret key, which can be a number, a word, or just a string of random letters, is applied to the text of a message to change the content in a particular way. This might be as simple as shifting each letter by a number of places in the alphabet. As long as both sender and recipient know the secret key, they can encrypt and decrypt all messages that use this key.

The problem with secret keys is exchanging them over the Internet or a large network while preventing them from falling into the wrong hands. Anyone who knows the secret key can decrypt the message. One answer is asymmetric encryption, in which there are two related keys--a key

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Symmetric key encryption is a form of cryptosystem in which encryption and decryption are performed using the same key. It is also known as conventional encryption.

Asymmetric encryption is a form of cryptosystem in which encryption and decryption are performed using the different keys – one a public key and one a private key. It is also known as public-key encryption [3].

A Key is a numeric or alpha numeric text or may be a special symbol. The Key is used at the time of encryption takes place on the Plain Text and at the time of decryption takes place on the Cipher Text. The selection of key in Cryptography is very important since the security of encryption algorithm depends directly on it. The strength of the encryption algorithm relies on the secrecy of the key, length of the key, the initialization vector, and how they all work together.

Asymmetric encryption techniques are about 1000 times slower than Symmetric encryption which makes it impractical when trying to encrypt large amounts of data. Also to get the same security strength as symmetric, asymmetric must use a stronger key than symmetric encryption technique. pair. A public key is made freely available to anyone who might want to send you a message. A second, private key is kept secret, so that only you know it.

II. Analysis of symmetric encryption algorithms

Symmetrical encryption algorithm has the key used to encrypt messages that can be obtained from the decryption key and vice versa [2].

In symmetric algorithms, legal user P by means of cipher device Cn turns sequence X = (x1, …, xn), which is called the public information, into the encrypted data Y = Cn(x) (Fig. 1).

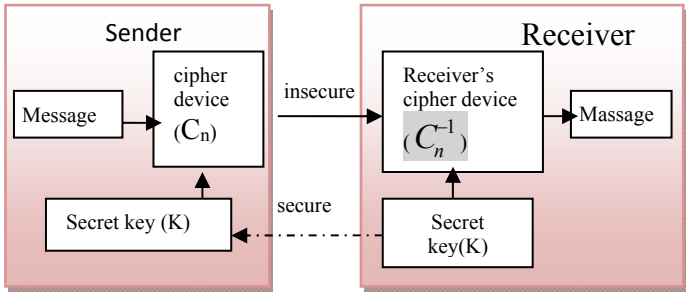


Fig. 1. Structure of the symmetric encryption scheme

The algorithm of the cipher device Cn depends on the parameter K = KX (Key), a known user. Legal users, who possess the information X, perform decryption of information using an algorithm that depends on a parameter K' associated with K. Usually, $K' = K$. In this case, every legal user who originally owns a transformation Cn, and transformation C_n^{-1} - reverse Cn, while the illegal user does not have the key K, which is not fully aware of the conversion Cn and C_n^{-1} [4].

Symmetric cryptosystems are based on the flow and block encryption algorithms. In the flow algorithm, every bit of plaintext is encrypted (and decrypted) by adding module 2 with bit of pseudo-random sequence – cryptographic bit stream, independently of the other bits. Thus, transformation of each intext symbol changes from one symbol to another [5]. Stability of flow encryption algorithm depends on whether the derivative has the property of equal occasional occurrence of the next symbol.

The advantages of streaming algorithms are the high encryption speed, relative simplicity, and the absence error propagation.

The disadvantages are:

– cryptographic bit stream shall not be used more than once (in terms of safety);

– the requirement of operations synchronicity at transmitter and receiver, which is expressed in the transmission timing of a random sequence in front of the message header before its decryption (so-called pseudo-random additional key, which is used to modify the encryption key for improving cryptographic robustness).

A plain text is first partitioned into blocks of equal length for block encryption algorithms, and then is ciphered within each block function depending on the key block into encryption text of the same length [5]. In the case where the length of the plaintext is not aliquant to input block length, multiple encryption algorithm shall be used to supplement operation of the last block of plaintext to the desired length. The essence of the block cipher algorithm is repeatedly applied to the plaintext block of mathematical transformation so as to set a dependency of each bit from the ciphertext and the plaintext key. Block

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

algorithm shall be designed in such a way that the change of even one bit of the plaintext and the key would result in a change of approximately 50% ciphertext bits, while none of the plaintext bit should never be administered directly into the ciphertext [3]. Conversion algorithms based on the data, are divided into a complicated (nonlinear operation) and simple (which are based on mixing), while the first construction provides cryptographic robustness. The most common block encryption algorithms:

1) mode of simple replacement and codebook mode (identical plaintext blocks are encrypted in the same way by the same key);

2) counter mode (initial state is defined by the original range of synchronous communication link, received gamma is processed through block encryption algorithm and then summed in module 2 with the plaintext block);

3) output counter mode (the same synchronous communication and feedback available on the ciphertext, counter mode is performed before the resulting unit will be converted by block encryption algorithm).

The advantages of block encryption algorithms (other than simple replacement mode) are:

Each ciphertext bit depends on all the bytes of the plaintext block and no two plaintext blocks are not represented by the same ciphertext block;

The possibility of application of such algorithms to detect manipulation of the messages made by meddlers.

It uses the fact of error propagation in ciphers and the ability of systems to easily generate a message authentication code.

Disadvantages of block encryption algorithms:

subject to restrictions of cryptanalysis "using the dictionary";

connected with reproduction error (as one error bit in transmission can cause a number of errors in the decrypted text);

development and implementation is more difficult than streaming encryption systems have.

In practice, long messages encryption is applied at inline block algorithms or algorithms with feedback. Repeated alternation of simple permutations and substitutions, managed by a long enough secret key, provides a fairly stable block algorithm with good dispersion and mixing. [3]

The most popular nowadays symmetric encryption algorithms are distinguished: DES, IDEA, GOST 28147-89, Triple, RC2, RC5, BLOWFISH and others.

Each symmetric algorithm is evaluated on the following criteria:

- dimensions of the input and output units;
- key size;
- complexity of data conversion algorithm;
- speed data conversion and cryptoattack resistance.

Stability data rate and conversion was evaluated on 6-level scale (6 - minimum, 1 - maximum [1] (Table 1).

Algorithm	RC5	FEAL	BLOWFISH	ГОСТ 28 147-89	IDEA	DES
Input block size, bit	32,64 or128	64	64	64	64	64
Output block size, bit	32,64 or128	64	64	64	64	64
Key size, bit	from 0 to 2040	64	448	256	128	56
Number of conversion cycles in algorithm	from 0 to 255	From4 to 32	16	32	12	16
Persistence of algorithm	6	4	2	1	5	3
Conversion speed	1	4	3	6	2	5

Table 1, Results of the comparison of symmetric encryption algorithms

Persistence of symmetric encryption algorithms is considered through the following criteria:

- key size;
- complexity of data transformation;
- existence of an algorithm.

In the viewpoint of cryptanalysis, the existence of an algorithm plays an important role. If the algorithm is used for a long time, it becomes an attractive target for cryptanalysts [3] and significant computing resources can be allocated to disclose the encryption algorithm. A DES algorithm can be a famous example of such an algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

According to Table 1, the most resistant to cryptoattacks of enemy is an encryption algorithm GOST 28147-89, but it is considered to be the slowest.

Modern information systems may use symmetric encryption methods in order to prevent unauthorized access to information in the absence of the owner. It can be both an archive encryption of selected files, and automatic encryption of entire logical and physical disks.

Symmetric algorithms are also used to protect data transmitted over open communication channels. [5]

The study of Asymmetric cryptosystems

The essence of the public key or asymmetric cryptosystems of two interrelated keys by a certain rule is generated by each addressee. [4]

The encryption public key scheme is shown in Figure 2.

One key is used for data encryption, the other – for decryption. Each of the correspondents has a key $k = (k_з, k_p)$ consisting of an public key $k_з$ and private key k_p . The open key encryption rule defines Ek, a secret key - decryption rule Dk. These rules are related (for any plaintext X and any ciphertext Y):

$$D_k(E_k(X)) = Y$$

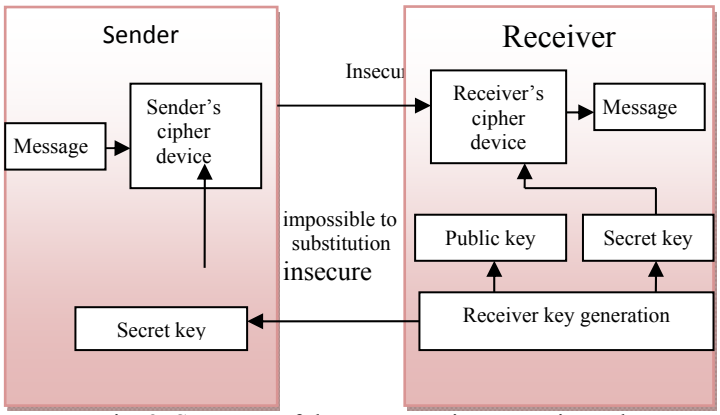


Fig. 2. Structure of the asymmetric encryption scheme

The knowledge of the public key does not allow determining a secret key in a reasonable time (or with reasonable complexity). Let state encryption and decryption rules (on selected key k) of arbitrary correspondent A by EA and DA symbols, respectively. Correspondent B wants to send a private message X to correspondent A, receives a copy EA, calculates the ciphertext Y = EA(X), which directs by communication channel to correspondent A. Correspondent B received message Y, applies DA conversion, receiving plaintext X.

Cryptographic public key systems use irrevocable or unilateral functions that have the following features: given value of X it is relatively easy to calculate the value $f(x)$, but if $y = f(x)$, there is no easy way to calculate the value of X. In other words, it is very difficult to calculate the value of the inverse function $f^{-1}(y)$ [3]. The study of irreversible functions is carried out mainly in three areas: discrete exponentiation; multiplication of prime numbers; combinatorial problems, especially the problem of concluding a portfolio.

Comparison of asymmetric cryptosystems is conducted according to the following criteria: the speed of used algorithms and the mathematical transformation of the information. The data conversion was evaluated on a 5-point scale (1-highest, 5-lowest score). Results of asymmetric encryption comparison techniques are shown in Table 2.

Algorithm	Conversion	Speed
RSA	discrete exponentiation, expansion of factoring	5
Diffie-Hellman	discrete exponentiation	2
El-Gamal	discrete exponentiation	3
Massey Omura	discrete exponentiation	4
Knapsack system	Problem backpack stacking system	1

Table 2.Results of the comparison of asymmetric encryption methods

RSA is considered the most persistent of the existing algorithms, since it is only once failed to disclose RSA cipher for 500-digit key. For these purposes, in 1600 computers of volunteers have been involved in within 5 months of continuous operation [1]. It should be noted that using the RSA system with keys 512-1024 bits is practically impossible to break ciphers. However, RSA system operates in thousand times slower than DES algorithm, and requires that the keys to be approximately 10 times longer.

While it is clear that the use of public key systems can be limited by challenge key exchange, followed by their use in symmetric cryptography that is the use of so-called hybrid systems [4]. The results of the comparison of classical cryptographic algorithm DES and cryptographic algorithm RSA with public key are shown in Table 3.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Characteristic	DES	RSA
Speed	Fast	Slow
Function used	Permutation and substitution	Involution
Length of the key	56 bit	300…600 bit
Least expensive cryptanalysis	Iterate over the key space	Module decomposition
Temporary costs on cryptanalysis	Centuries	Depends on the key length
Key generation time	Millisecond	Tens of seconds
Type of key	Symmetric	Asymmetric

Table 3 Results of comparing DES and RSA algorithms

When analyzing the strengths and weaknesses of symmetric and asymmetric systems, it is determined that the asymmetrical encryption systems provide a significantly lower encryption rate than symmetrical, that is why they are usually used not only to encrypt messages, but as encryption of keys exchanged between correspondents, which are then used in symmetric systems.

The main advantage of public key cryptosystems is their potentially high safety: there is no need to transfer or disclose to anyone the value of the secret key, to make sure of their reliability. In symmetric cryptosystems, there is the risk of disclosure of the secret key during the transmission.

However, the algorithms that base public key cryptosystems have the following disadvantages:

- Generation of new private and public keys based on the new generation of large prime numbers and primality testing takes a lot of device time;
- encryption and decryption processes are related the construction of the power of a multi-valued number, are rather cumbersome.

Therefore, the speed of public key cryptosystems is usually hundreds times or even more less than the speed of symmetric sector key cryptosystems.

Asymmetric encryption algorithms are used to solve many problems: user authentication and message, generation of session keys in information systems, systems for identification “friend-or-foe”.

III. Conclusions

- The studies of modern methods and algorithms for cryptographic protection of information from unauthorized access can conclude that modern information systems for the encryption of transmitted messages use symmetric encryption algorithms. Asymmetric algorithms, because of their large computational complexity, are used for the generation and propagation of session keys.
- The combined use of symmetric and asymmetric encryption allows eliminating the main drawbacks of both methods.
- The combined method of encryption keeps the advantages of high security provided by asymmetric cryptosystems with a public key, and the advantages of high speed operation, inherent in symmetric cryptosystems with a secret key. The proposed approach allows choosing the method of protection based on the selected performance criteria, as well as assessing the possibility of practical use of the considered cryptographic methods of protection.

References

[1]. Burnett S., Payne S. RSA Security's Official Guide to Cryptography. – M.: Bean, 2002. – 382 p.

[2]. Rostovtsev A.G., Mahovenko E.B. Theoretical cryptography. – M.: Publ. Professional, 2005. – 490 p.

[3]. Salomaa A. Public key cryptography. M.: Mir, 1995. — 318 p.

[4]. Schneier B. Applied Cryptography: Protocols, Algorithms, Source codes in C language. — M.: Triumph, 2002. — 816 p.

[5]. Blahut R.E. Cryptography and Secure Communication. – Cambridge University Press, 2014. – 608 p.

[6]. Hazem Hatamleh (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, 2016

SQS: An Ontology Based Framework for Evaluating Service Oriented Software Applications: A case study of E-Governance Services

Ezekiel U. Okike
Department of Computer Science,
University of Botswana, Gaborone, Botswana
euokike@gmail.com

***Abstract*—E-governance is about enabling good governance through the use of modern Information and Communication Technology. As a service based ICT platform, the main challenge is efficient and effective evaluation framework. In this paper, an ontology based framework for evaluating e-government software applications is proposed. The proposed framework uses a three stage model: standardization, quality, and service stages (SQS). The model provides effective and dependable evaluation of e–governance from users and stakeholders' perspectives.**

Keywords-ontology; e-governance; framework; evaluation model .

I. INTRODUCTION

A Ontology is a formal approach to specifying a concept and its representation of a domain [1]. The concept being represented is explicitly described using formalisms or other appropriate representation that provides a description of the concepts and the relations between them as well as its technological components [2,3].

Using ontologies computational models are created for automated reasoning in Artificial intelligence [4]; classes, relations, functions and objects are defined in Object Oriented systems [5]; common understanding of objects are shared; knowledge reuse, and explicit assumption are enabled; and domains are separated and analyzed [6,4]. Furthermore, ontologies are used in classifying object based on scope or domain granularity, taxonomy construction direction, and the type of data sources [7]. Figure1 shows the various levels of ontology classification: the base level (application level), the intermediate level (domain oriented and task oriented), and the top level.

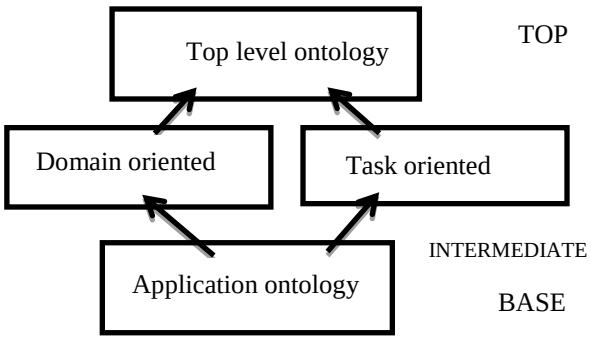


Fig. 1. Ontology classification
Source: Adapted from Antonio [8].

In this model, the top-level ontology has some concepts that have general agreements or stable standards. Domain ontology has concepts that define the main focus of interest on the domain. Task ontology deals with sub-concepts that are needed to solve problems on the main domain and the application ontology deals with concepts that exercise the fastest rate of exchanging data [8].

A. Problem Statement

The search for effective and efficient evaluation model of e-government services continues as current evaluation option are still evolving especially in developing countries. This is so because in the developing countries the concept of e-governance is still poorly implemented and lacks appropriate standards. Hence, the need for efficient evaluation framework cannot be over emphasized.

B. Objectives

The main objective of this paper is to study existing e-government models and hence propose appropriate framework for evaluating e-government software services.

This rest of this paper counts of three sections. Section 2 is a review of the literature on e-government concept and how to build an ontology for e-government. Section 3 looks at

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ontology based e-government models. Section 4 proposes an evaluation model for e-government, while section 5 is conclusion and future work.

II. LITERATURE REVIEW

A. The Concept of E-governance

E-governance is about enabling good governance through the use of modern Information and Communication Technology (ICT). The concept (also known as Digital governance) implies the growing use of ICT as a catalyst for the formation of knowledge societies where people have more access to relevant information as participants in their own governance and development. According Nath [9], "Knowledge networks function on the underlying principle that access to information is empowering and strategic use of information by citizens could become the key to popular and meaningful governance". This assertion is premised on the knowledge networking model shown in figure 2.

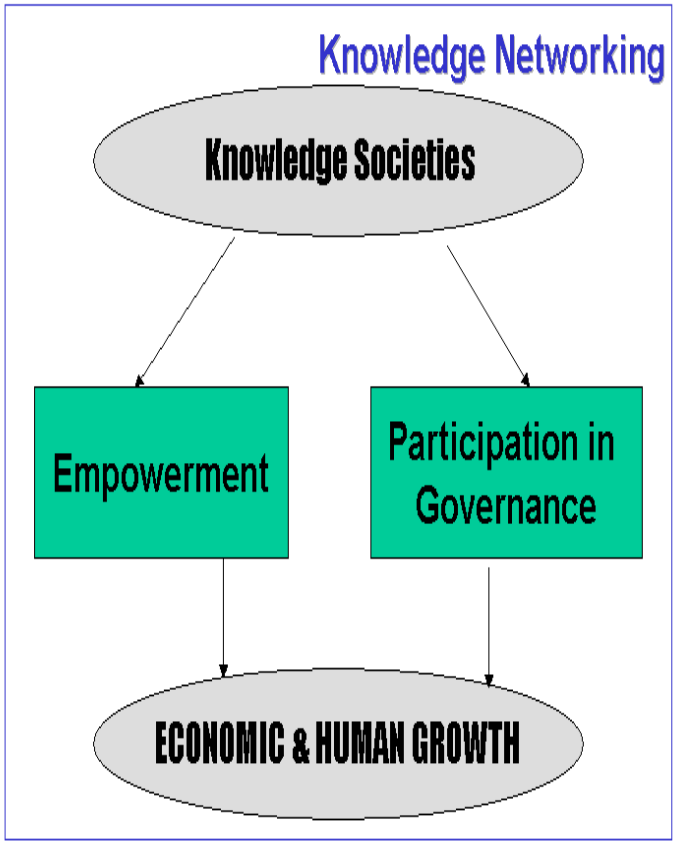


Figure 2: Knowledge Networking through ICT empowerment (Source: Nath, V [9])

Although, e-governance (digital governance) is still evolving in developing countries, there are five generic models in use [9]. These include Broadcasting/wider dissemination model, Critical Flow model, Comparative Analysis model, Mobilization and Lobbying model, and Interactive–service model. The underlying principle, applications and organization of each model is summarised as follows:

i) Broadcasting/Wider-Dissemination Model – This model aims to disseminate information for better governance through the use of ICT. This helps the citizenry understand governance so that they are able to make informed decisions.

ii) Critical Flow model - This model aims to channel information of critical value to targeted audience through the use of ICT. Using ICT such information is disseminated timely irrespective of distance.

iii) Comparative Analysis model- This model aims to explore information available in the public or private domain, and compares that with already known information for strategic purposes. Therefore, new and assimilated information are used as benchmark for governmental advocacy and policies

iv) Mobilisation/Lobbying model- This is a digital governance model often used by civil society organizations in order to make their influences and impacts known through virtual communities.

v) Interactive-service model – this model aims to offer government services to the citizens using interactive ICT channels such as e-voting, e-tax, e-procurement e.t.c

B. Building an E-governance Ontology

Several approaches could be followed to build an ontology based e-governance. One could either use the bottom–up approach, top-down approach or the middle out approach (Catherine Roussey et al, 2011).

Bottom-Up approach: defines first the most general concept of the entity in use then goes towards the most specific aspects.

Top-Down approach*:* defines first the most specific concepts then goes towards the most general aspects.

Middle-Out approach*:* defines the concepts from the central area towards the general and / or specific concepts. Therefore, an e-government ontology may be defined following these principles.

According to Roussey [7], ontologies could also be described according to sources used to get the knowledge. The knowledge could either be based on:

Text*:* Unstructured data given to a computer system for processing.

Thesaurus*:* forming concepts from words or linguistic relations to build ontology.

Relational Database*:* structured and accurate software storages used to build ontologies from.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

UML Diagrams*:* using formal described UML classes to define concepts to build ontologies

In addition, an e-governance ontology can also be defined using the Enterprise Ontology Modelling Process (EOMP) identified by Uschold and Gruninger [10]. Using this approach requires the following:

i) Identify Purpose and Scope: which deals with main reason why the ontology is being built

ii) Building the ontology**:** segmented into three steps

(a) Ontology capture*:* deals with identifying the key concepts and relationships in the domain of interest.

(b) Ontology coding*:* deals with representation of the knowledge using a formal language for the ontology.

(c) Integrating existing ontologies*:* incorporates the both coding and capturing process with logic of how to use the ontology.

(d) Evaluation: gives a technical judgment on the ontology

(e) Documentation: Stating the guidelines for each purpose

Furthermore, ontology development process could be done following the IEEE standard for developing Software Life Cycle Process [11].

III. ONTOLOGY BASED E-GOVRNANCE EVALUATION MODEL

E-governance is a software based online/web based service. Hence, some of the principles of measurement in software are very useful in evaluating e-governance structures.

A. Ontologies in Software measurement

Generally, measurement is a mapping from the empirical world to the formal, relational world**.** Consequently, a measure is the number or symbol assigned to an entity by this mapping in order to characterize an attribute [12].

Theoretically, Measurement Theory (MT) species the rules for developing and reasoning about all kinds of measurement. As explained in [14], rule based approach is common in the sciences such as Chemistry, Physics and Mathematics. In Mathematics, Mathematicians learned about the world by defining axioms for a geometry. Hence, by combining axioms and using their result to support or refute their observations, they expanded their understanding and the set of rules that govern the behavior of objects.

In any software measurement activity the entities and attributes to be measured must be clearly identified and specified.

In software measurement, three software activities are involved namely:

i) Processes – collections of software related activities

ii) Products - artefacts, deliverables or documents resulting from process activities

iii) Resources – entities required by a process activity

Software artefacts have 2 essential types of attributes namely internal and external attributes.

Internal attributes are measured in terms of the product itself. Essentially, internal attributes are code based measure of software quality attributes such as class cohesion, class coupling, control structures, algorithms, data structures, and nesting level[13].

External attributes are measured in terms of how the software product, process or resource relate to the environment of operation. The measures are aimed at evaluating the software from the users perspectives in terms of its usability, reliability, efficiency, reusability, maintainability, portability, and testability e,tc. Figure 3 below shows the standard ISO/EC 9128 evaluation guide based on external software attributes. This guide is a useful ontology based model for all aspects of internal and external software quality measures.

External attributes (figure 2) are measured in terms of how the software product, process or resource relate to the environment of operation. The measures are aimed at evaluating the software from the users perspectives in terms of its usability, reliability, efficiency, reusability, maintainability, portability, testability e.tc. ISO 9126 [15] proposed a standard which species six areas of importance, i.e. quality factors, for measuring external software attributes. These include functionality, reliability, efficiency, maintainability, portability, and usability. This model was has since evolved into the ISO/EC 9128 [16] software product evaluation standard as shown in figure 2.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

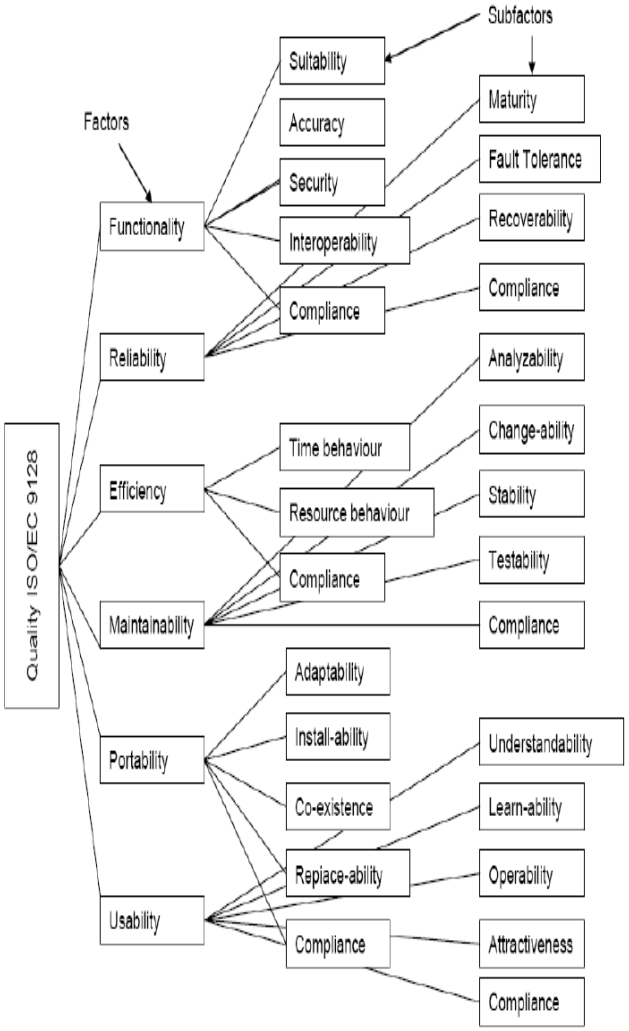


Fig. 2: ISO/EC 9128: Software Product Evaluation: Quality Characteristics and Guidelines for their Use.

This model has evolved into ISO/EC 25010. A detailed review of software quality models for the evaluation of software products is presented in Miguel, Mauricio, and Rodriguez [17]. However, in this paper, although all the models are useful, the ISO/EC 9128 standard is used. By integrating standard e-governance model, and the ISO/EC 9128 or ISO/EC 25010 this paper proposes an evaluation framework for e-governance as described below.

IV. SQS:AN ONTOLOGY BASED FRAMEWORK FOR EVALUATING E-GOVERNANCE.

SQS is an acronym for Standards, Quality and Service. Hence, the proposed e-governance framework is focused on the following aspects:

i) Standards. Any e-governance evaluation should begin by ascertaining if the e-governance in place is modelled after acceptable e-governance standard such the one defined by
   a) The broadcasting model
   b) The critical flow model
   c) The organisation/project based model
   d) The comparative analysis model
   e) The mobilization and lobbying model
   f) The interactive service model

The key question to answer is "Does existing e-governance follow acceptable standard?" i.e. does it take care of the items "a-f" in its implementation ?.

ii) Quality (Quality of Service QoS). The QoS of a Service Oriented Software Initiative (SOSI) such as e-governance will be better evaluated using both the internal and the external software quality attributes such as the standard ISO/EC 9128-ISO/EC 25010 software product evaluation quality characteristics. This can be done by designing appropriate questionnaires which capture all desired external attributes for the users of the e-governance service. By analysing collected feedback, and interpreting results, a good evaluation of any e-governance service may be obtained in terms of its QoS based on the factors identified in figure 2.

iii) Service Delivery
A Service Delivery Framework (SDF) is a set of principles, standards, policies and constraints to be used to guide the design, development, deployment, operations and retirement of services delivered by a service provider with a view to offering consistent service experience to a specific user community in a specific community. The important question to answer in evaluating an e-governance is "Is there any Service Delivery (SDF) model in place? This implies ascertaining that principles, policies, standards and constraints in respect the existing e-governance are in place. If these are in place, the next question to answer is "Is service delivered ?" "By what indicators?" Measurable indicators of service delivered could be achieved by :
   a) Specifying expected output indicators
   b) Ascertaining service effectiveness
   c) Ascertaining user satisfaction
   d) Ascertaining service availability
   e) Ascertaining service functionality
   f) Ascertaining service reliability
   g) Ascertaining service measurability
   h) Ascertaining service accountability

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

i) Ascertaining service manageability e.t.c

Outcomes are the end result that the government wishes to achieve through its e-governance initiative, and in particular with reference to how the rural populace benefit from the e-governance service. Indicators assess the impact of the program output on the desired outcomes that government want to achieve in the e-governance initiative.

A. *Measuring Service delivery*

The following steps are necessary in order to measure service delivery:

a) Clarify service delivery and performance measurement tools
b) Specify appropriate measureable objectives and output
c) Develop robust output measures and indicators.

B. *Relationship between internal and external Attributes*

Internal software attributes are code level measures of the quality of the underlying codes of the software. Some code level measure include cohesion, coupling, lines of code, cyclomatic complexity, Depth of inheritance

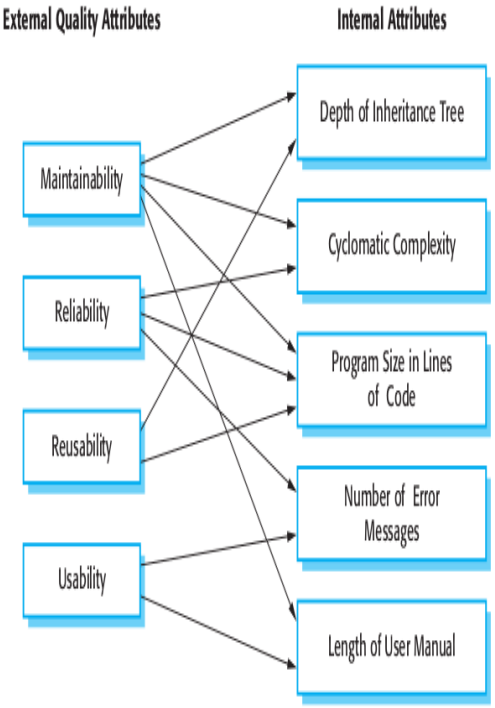


Figure 4. Relationship between internal and external quality attributes

For evaluating e-governance, using internal attributes of the software are not recommended, but using external attributes are highly recommended. This is because rural users are the object of measuring the success of e-governance initiatives.

V. CONCLUSION

As a service oriented software platform, e-government success hinges on service delivery. Successful service delivery models are based on appropriate standards, and policies which are also part of the software implementation.

In this paper, a three stage model for evaluating e-governance has been proposed. The stages in this evaluation model include standardization, quality and service (SQS) . The future direction of research on this paper will focus on empirical studies based on the SQS framework.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] Kang Ye etal. (2009). Ontologies for crisis contagion management in financial institutions. *Journal of Information Science, 35 (5)*, 548–562.

[2] Silvonen, P. (2002, October 21). *Ontologies and Knowledge Base*. Retrieved October 22, 2015, from http://www.ling.helsinki.fi: http://www.ling.helsinki.fi/~stviitan/documents/Ontologies_and_KB/ontology.html

[3] Sunitha Abburu and G Suresh Babu. (2013). A Framework for Ontology Based Knowledge Management. *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-3*, 21-25.

[4] Gruber, T. (2009). Ontology. In L. L. Özsu (Ed.), *Encyclopedia of Database Systems.* Springer-Verlag.

[5] Gruber, T. R (1993). A Translation Approach to Probable Ontology Specification. Knowledge Acquisition, 5(2):199-220

[6] Kabilan, V. (2007). *Ontology for Information Systems (O4IS) Design Methodology: Conceptualizing, Designing and Representing Domain Ontologies.* The Royal Institute of Technology: KTH information and Communication Technology .

[7] Catherine Roussey et al. (2011). An Introduction to Ontologies and Ontology Engineering. In F. G. al, *Ontologies In Urban Development Projects* (p. 241). London: Springer-Verlag.

[8] Antonio Zilli et al. (2009). *Semantic Knowledge Management: An Ontology-Based Framework: An Ontology...* New York: Information Science Reference

[9] Nath, V (2007). Digital Governance Models: Towards Empowerment and Good Governance in Developing Countries in Rafael Capurro, Johannes Frühbauer, Thomas Hausmanninger (eds). Localizing the Internet. Ethical Aspects in Intercultural Perspective. Munich . ISBN 978-3-7705-4200-0

[10] Mike Uschold and Micheal Gruninger. (1996). Ontologies: Principles, Mthods and Applications. *Knowledge Engineering Review Volume 11 No 2* , 1-69.

[11] Mohammad Nazir Ahmad et al. (2012). *Ontology-Based Applications for Enterprise Systems and Knowledge Management.* United States of America: IGI Global.

[12 Fenton, Norman and Pfleedger, S. L (1997). Software Metrics: A Rigorous and Pratctical Approach, 2nd ed, Boston, M.A:PSW Publishing

[13] Ezekiel U. Okike, and Tebo Leburu (2017). Axiomatic and Measurement Theory based Framework for Selecting Success Factors in Software Project Management. International Journal of Computer Science and Network Security, Vol. 17 No. 12 , pp.

[14] Okike, Ezekiel U (2007). Measuring Class cohesion in Object oriented Systems using Chidember and Kemerer Metric Suite and Java as case study. PhD Thesis Department of Computer Science, University of Ibadan. Unpublished

[15]ISO/EC9126.http:/www.bth.se/com/besq.nsf/(webfiles)/…/$FILE/chapter_1.pdf

[16] ISO/EC9128http:/www.bth.se/com/besq.nsf/(webfiles)/…/$FILE/chapter_1.pdf

[17] Miguel, Jose P; Mauricio, D; and Rodriguez, G (2014). A review of Software Quality Models for the Evaluation of Software Products. International Journal of Software Engineering & Applications (IJESA), Vol. 5, No. 6, pp. 31-53.

AUTHORS PROFILE

Dr Ezekiel Uzor Okike, MIEEE, MACM, MAIS is a Senior Lecturer in the Department of Computer Science, University of Botswana, Gaborone. He holds a PhD in Computer Science (2007), a Master of Information Science (M.Inf,Sc), 1994 and a B.Sc Computer Science, (1992) all from the University of Ibadan, Nigeria. He is a member of IEEE, ACM, and AIS. Presently he is the cluster chair of the Information Systems Cluster, Department of Computer Science, University of Botswana. He has published many papers in international journals and attended many international conferences where he presented papers. His research interests are in Software Engineering, Software Measurement and Models, Software Architectures, Information Systems, Cyber Security, and High Performance computing.

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2017-2018
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/